REFINEMENT OF REDUCED PROTEIN MODELS

WITH ALL-ATOM FORCE FIELDS

A Dissertation
Presented to
The Academic Faculty

by

Liliana Wróblewska

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Biology

Georgia Institute of Technology
December 2007

REFINEMENT OF REDUCED PROTEIN MODELS

WITH ALL-ATOM FORCE FIELDS

Approved by:

Dr. Jeffrey Skolnick, Advisor
School of Biology
Georgia Institute of Technology

Dr. John McDonald
School of Biology
Georgia Institute of Technology

Dr. Facundo M. Fernandez
School of Chemistry & Biochemistry
Georgia Institute of Technology

Dr. David C. Sherrill
School of Chemistry & Biochemistry
Georgia Institute of Technology

Dr. King Jordan
School of Biology
Georgia Institute of Technology

Date Approved: November 13, 2007

To my Parents who have always reached further than they could see

Moim Rodzicom, którzy zawsze sięgali dalej niż byli w stanie ogarnąć

ACKNOWLEDGEMENTS

I wish to thank my friends and family for their support, guidance and motivation. My dear partner and best friend, Piotr Rotkiewicz, for always being there for me with unwavering kindness, patience and ideas. My best collaborator and best mate, Anna Jagielska, for all the things we have done together and we will continue to do in the future. My sister and brothers with their families, for the greatest pleasure of growing with you and next to you. All of you are always my best motivation for work and development.

Secondly, I want to thank all the past and present members of the Skolnick group. For your help and comments, for the good words and a great atmosphere in the lab. I also want to thank my former collaborators from the Quantum Chemistry Laboratory at Warsaw and the Hauptman-Woodward Institute at Buffalo for all I have learned from you.

Finally, the one without whom none of these would have been possible, my advisor Jeffrey Skolnick. I admire his passion for work, his focused drive to get where he is. Thank you for taking me in your group and believing in my abilities. Thank you for your encouragement and perseverance.

I would also like to thank my committee members Dr. Facundo M. Fernandez, Dr. King Jordan, Dr. John McDonald and Dr. David C. Sherrill for their great inputs, help and patience.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF SYMBOLS

Å	angstrom
α	alpha
β	beta

LIST OF ABBREVIATIONS

AMBER	Assisted Model Building with Energy Refinement
CAS	C-Alpha + Side chain
CC	Correlation Coefficient
DIH	dihedral energy
DSSP	Dictionary of Secondary Structure of Proteins
ELE	electrostatic energy
ELE1-4	electrostatic energy for atoms separated by less than four bonds
ff	force field
GB	Generalized Born
MC	Monte Carlo
MD	Molecular Dynamics
ns	nanosecond
NMR	Nuclear Magnetic Resonance
PDB	Protein Data Bank
ps	picosesond
RMSD	Root Mean Square Deviation
SICHO	Side Chain Only
SA	Surface Area
TASSER	Threading/ASSEmbly/Refinement
VDW	Van der Waals energy
VDW1-4	Van der Waals for atom pairs separated by less than four bonds

SUMMARY

The goal of the following thesis research was to develop a systematic approach for the refinement of low-resolution protein models, as a part of the protein structure prediction procedure. Genome sequencing projects are producing amino acid sequences, but a full understanding of the biological role of these proteins will require knowledge of their structure and function. Although experimental structure determination methods provide high accuracy atomic coordinates for a subset of proteins, experiments are still too costly and time consuming to be used on a genomic scale. Computational structure prediction methods provide an appealing alternative for fast, large-scale structure determination. Significant progress has been made in the field [1-3] and the contemporary protein structure prediction methods are able to assemble correct topology for a large fraction of protein domains. But even such approximately correct models typically vary in structural similarity to the native structure from 1 to about 6 Å RMSD (root mean square deviation). Models with a resolution of 1-2 Å have a reliability comparable to the experimentally obtained structures and can be used in a broad range of applications, including studies of reaction mechanisms, functional annotation, drug design, virtual ligand screening and others. For low-resolution models (3 - 6 Å away from the native) the spectrum of useful applications is much narrower [2]. Unfortunately, the problem of protein model refinement has seen little success so far.

According to the thermodynamic hypothesis [4], the native state of a protein is a free energy minimum conformation of a given polypeptide chain. Also, in order for the protein folding to be efficient, the free energy surface has to have somewhat a funnel-like

shape [5,6] that guides the chain toward its native state. The theoretical free energy surface should also exhibit similar characteristics. In the real protein prediction, a search method is used to generate alternative conformations of a given protein chain and the structures are subsequently judged with a theoretical free energy function. The minimal requirements for a potential that can refine protein structures is the existence of a correlation between the energy with native similarity and the scoring of the native structure as being lowest in energy. Only then the lowest energy conformations can be at the same time the best structural models for a given sequence.

The initial studies undertaken in this project were aimed at a systematic assessment of the existing state of the art all-atom protein simulation methods in the task of the model refinement. Following the thermodynamic hypothesis, we tried to answer two key questions about the characteristics of a given force field: 1) is the native structure the global free energy conformation of the tested potential? 2) does the free energy exhibit a correlation with native similarity; that is, is the potential able to drive the conformational search towards native-like structures? The results of our analysis revealed that commonly used all-atom potentials exhibit significant issues when the global shape is considered. Often the lowest energy structure is very far away from the native in the sense of RMSD, and there is no correlation between native similarity and energy. Consistently with these findings, during the conformational search driven by such force fields the majority of protein models drift farther away from the native. The clear conclusion from the test was, that the force fields needed to be corrected to be able to refine protein models.

In the second part of this project, we employed a large set of structural and energetic data, and a global optimization method, to reshape the potential function and make it funnel-like. We changed the relative weights of particular components of the tested all-atom force field in such a way that the final energy function has a global minimum in the native state and an improved correlation with native similarity. Optimization was conducted for a set of representative native protein structures and their decoys that span a wide range of similarity to the native. Such a global funnel-shaping approach proved to be a powerful method to significantly improve both native scoring and the correlation coefficient in the newly optimized potential. Additional improvement was made possible by adding an explicit formula for the hydrogen bond potential to the original force field.

The last part of the research focused on protein model refinement using the newly developed energy function. We performed conformational search driven by the optimized energy function, starting from a large set of protein models with varying native similarity. The test employed 47 proteins and 100 decoy structures per protein. When the lowest energy structure from each trajectory was compared with the starting decoy, we observed structural improvement for 70% of the models on average. Such an unprecedented result of a systematic refinement is extremely promising in the context of high-resolution structure prediction.

CHAPTER 1

INTRODUCTION

1.1 Protein Structure Prediction

Christian B. Anfinsen in his 1972 Nobel prize acceptance lecture [4] stated that: "The amino acid sequences of polypeptide chains (…) only make functional sense when they are in the three dimensional arrangement that characterizes them in the native protein structure". This is the summary of the protein sequence-to-structure-to-function paradigm and a motivation for all the protein structure determination studies. Experimental methods for structure determination include crystallography, nuclear magnetic resonance spectroscopy (NMR), and electron cryomicroscopy (cryo-EM). Crystallography provides often high quality protein structures that are used to elucidate reaction mechanisms, or the mode of binding between proteins and other molecules. Among the drawbacks of the method are problems with crystallization, which is very difficult for some proteins (e.g. membrane proteins), and the fact that crystallographic conditions are often very different from physiological conditions; that may influence the results. NMR studies are also very helpful in providing detail structural information, although the technique is applicable only to relatively small and soluble proteins. Cryo-EM elucidate only a general shape of the molecule, therefore are not useful for applications that require detailed structures. Furthermore, all experimental methods are costly and time consuming.

In the genomic era, the number of known protein sequences grows exponentially with time. On the contrary, the number of experimentally solved protein structures does

not even come close to the number of known sequences. Therefore, there is the need to fill the gap between the number of known sequences and structures. One possible means to achieve this goal is to employ computational methods for prediction of protein structure from sequence. All the contemporary protein structure prediction methods rely on the thermodynamic hypothesis and try to find the active conformation for the particular sequence, the native state, as the free energy minimum of the polypeptide chain.

1.2 Thermodynamic Hypothesis

In the same Nobel prize acceptance lecture [4] Anfinsen described the thermodynamic hypothesis of protein folding as follows: "The three-dimensional structure of a native protein in its normal physiological milieu (solvent, pH, ionic strength, presence of other compounds such as metal ions or prosthetic groups, temperature, and other) is the one in which the Gibbs free energy of the whole system is lowest; that is, that the native conformation is determined by the totality of interatomic interactions and hence by the amino acid sequence, in a given environment". The hypothesis was based on the observation that many proteins undergo a reversible denaturation, including disulfide bond rupture and reformation. Further studies reviewed by Kim and Baldwin [7] and Dill [8] confirmed the hypothesis giving evidence that for many proteins folding and unfolding reactions reach an apparent equilibrium. Although later experiments showed there are notable exceptions of proteins that fold under kinetic control [9], the majority of protein native structures are likely to be at global free energy minima for their amino acid sequences and physiological conditions.

With the thermodynamic hypothesis it is very easy to formulate an idealized protocol for the protein structure prediction. For a given polypeptide chain it would involve:

1) generation of all possible conformations of the chain,
2) calculation of a free energy for each of the generated configurations.

The lowest energy conformation would be the native state. Unfortunately neither of these two steps is possible to realize and the protocol is just an illustration of the two-fold challenge faced in the problem of structure prediction. As proteins have a vast number of degrees of freedom, and therefore an immense conformational space is available to them. Even if it is assumed that each amino acid populates only two regions of the Ramachandran plot [10], there are still on the order of 2^{100} ($\sim 10^{30}$) main-chain conformations for a small protein with 100 residues. With the speed of evaluation of e.g.1000 conformations per second it would take $\sim 10^{20}$ years to search all 10^{30} configurations. And still this is only a very crude approximation, since amino acids have many more then two degrees of freedom.

The energy calculation is equally challenging. The most accurate quantum methods and the contemporary computers can handle on the order of tens of atoms, while proteins are composed of thousands of atoms. Moreover, the time of energy calculation grows polynomially with the size of the system, on the order of N^5, where N is proportional to system size. The energy evaluation should also account for the protein's environment, e.g. solvent and ions, making the task even more complex. Therefore, considerable ingenuity has gone into the development of wide variety of methods for

reducing the size of the conformational space, sampling the space more efficiently, and simplifying the potentials.

1.3 Conformational Search

To overcome the intractable problem of the vast conformational space search, different approaches have been developed that focus mainly in two areas: simplifications of the polypeptide chain and reduction of the sampled conformational space by probing only in the important part of the space or lattice-based search.

An alternative to all-atom representation of a protein is a coarse-grained model in which a group of atoms is reduced to a single interaction center. Different coarse-grained models characterize proteins at varying levels of details. The minimalist model represents each amino acid with a single particle, such as side chain center of mass in the SICHO model [11], or with two particles - one for the backbone and one for the side chain. An example of the later approach is the CAS model used by *TASSER* [12,13], the protein structure prediction tool developed in our laboratory, where the interaction centers are placed at the C-alpha atom and at the side chain center of mass of each amino acid. Other used coarse-grained representations include the CABS model [14] (similar to CAS, but including one more particle placed at the C-beta atom), a model with all backbone heavy atoms and side chains represented by a single particle [15], or the united atom representation that treats a methyl group as a single interaction center [16,17]. Such models provide a large reduction in the number of degrees of freedom, and various algorithms have been developed to reconstruct the atomic details from the simplified representation with good accuracy [18,19].

Another important modification of the search is the reduction of the allowed conformational space prior to the energy calculation. This includes various means of freezing or restraining a part of the degrees of freedom. A good example is a template-based search where for a given protein sequence with unknown structure, sequence homologues with already solved structures are found. The well-aligned part of the protein chain is threaded onto the known scaffold and frozen or restrained during the conformational search [12,14,15]. A more general simplification is the use of any other information that can be inferred from already known structures, such as fixed bond lengths between interaction centers, or sampling only, or with higher priority, the most probable conformation (e.g. side chain rotamers most often found in protein structures). Also some of the already generated but wrong conformations can be rejected prior to the energy calculation based on simple judgments, such as excluded volume violations.

Yet another approach involves guiding the search into areas of low energy. This idea is employed in a variety of methods including stochastic methods, such as simulated annealing [20], Monte Carlo with minimization (MCM) [21], conformational space annealing (CSA) [22], or deterministic methods such as the diffusion – equation method (DEM) [23].

Finally, the search can be speed-up by discretizing the conformational space: the molecule is projected onto a specially designed lattice and during the search the interacting particles are "hopping" between the available lattice points [24,25].

1.4 Energy function

Conformational search in protein structure prediction is always driven by energy and it aims to find the global minimum of the free energy function. The particular functional form of the free energy depends on the molecular representation being used.

Quantum mechanics methods, although very accurate, consider the electronic structure of the system and the calculations are very time-consuming. Some of the largest molecules tackled by the quantum methods are peptides that include only 6-8 amino acids [26]. In the next widely used approximation, the electronic structure is ignored and atoms are considered as interacting particles. Such all-atom, physics-based force fields employ a simple model of interactions with contributions from the stretching of bonds, the opening and closing of bond angles, rotations about single bonds, and the non-bonded interactions such as van der Waals and electrostatic attraction or repulsion. Van der Waals interactions are often modeled by a Lennard-Jones type of potential and the electrostatic energy is calculated as a pair interaction of point charges placed at each atom. The force field parameters, the charges, the equilibrium values of bonds and angles, the force constants for their deviations, and others, are derived from experiment or quantum mechanical calculations. All these contributions comprise an approximation to the enthalpy of the protein molecule. A separate problem is the effect of solvation and entropy. In all-atom simulations, solvent (water) is often represented explicitly, but such an approach is extremely impractical for the use of protein structure prediction, since it greatly expands the search space. Therefore, simplified - or so called implicit or continuum solvent models - were developed. They are based on an approximation of the mean-force potential for the solvation interactions, that averages out the degrees of freedom of solvent molecules. The solvent is a virtual, infinite continuum medium with the dielectric and hydrophobic properties of water. The polar mean-field of the solvent polarization around the charged solute is approximated through the Poisson-Boltzmann [27] or the simpler, Generalized Born theory [28,29]. The non-polar contribution corresponds to

favorable van der Waals attractions between the solute and the solvent, and the unfavorable cost of breaking the structure of the solvent around the solute. This is usually modeled as being proportional to the solvent accessible surface area [30]. The polar and non-polar terms constitute the free energy of solvation. The last part of the total free energy, the entropy of the protein is most often neglected, since it is estimated that it is similar for the native and misfolded states [31].

Another class of force fields are knowledge-based potentials that are derived from statistics over the library of known protein structures. In this approach contributions to the total energy are dependent on the probability of the occurrence of certain instances in the real protein structure, e.g. a burial of a certain amino acid in the protein core or a close contact of a certain pair of amino acids in a particular orientation (parallel, antiparallel). The advantage is that the knowledge-based force field can be in principle derived for any of the simplified protein models. Also the knowledge-based potentials contain composite information about all the effects that contribute to the final structure of the native state (e.g. solvation). The disadvantage is the problem of statistics, that is the particular energy contribution has to be well represented in the database in order to be meaningful. Also each energy component in such potential is a result of multiple kinds of physical interactions; therefore, a set of independent contributions can only be obtained by extensive testing.

1.5 Hierarchical approach to protein structure prediction

The coarse-grained protein models proved to be extremely useful for simplification of the conformational search problem. But reduction of the complexity of the structure seriously compromises the resolution of the system at the same time, and

there is always a trade-off between the two. When fine atomic details of a molecule are present, the conformational space that can be searched in a realistic time is very limited. On the other hand, the protein representation suitable for an exhaustive search often lacks structural details that may be crucial, e.g. correct packing of the structural core. In effect, the simplifications in the protein representation influence the quality of the prediction and the applicability of the predicted model. The solution to this problem would be a hierarchical approach where the global conformational space search is performed in the first step for selection of approximate models and subsequently the all-atom model is reconstructed and refined with limited search in all-atom force field.

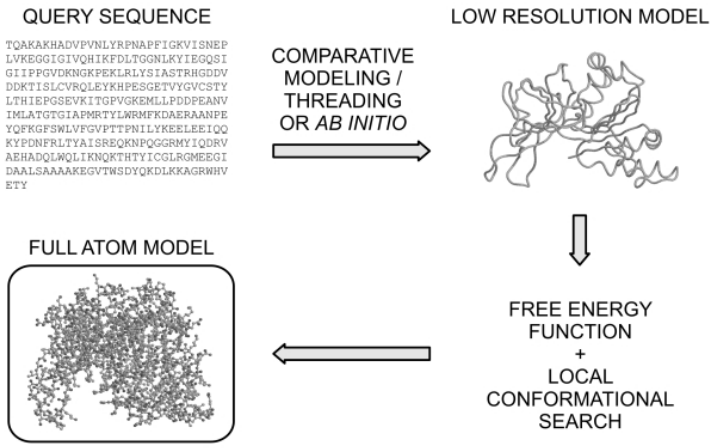


Figure 1.1 Schematic overview of the hierarchical approach to protein structure prediction.

1.6 Refinement of low-resolution protein models

Major attention of all the protein structure prediction related studies focused so far on the first step of the hierarchical approach, the development of simplified molecular models and force fields for effective conformational space search and the sensitive methods for finding relevant constraints that can reduce the search space. Over a decade ago, the CASP (Critical Assessment of protein Structure Prediction) competition has been launched [32] that gathers the structure prediction community every two years in the challenge of a blind prediction of protein structures that have been determined experimentally but not released to the public. The competition helps in continuous evaluation of the prediction methods, monitoring and boosting progress in the structure prediction area. The results of the contemporary state-of-the-art structure prediction methods are typically a set of models that range in quality from 1-6 Å root-mean-square-deviation, RMSD, of the backbone atoms from the native. Although such models have mostly correct topology, the key question is: can models of this quality be used for protein function prediction? There are a variety of function prediction methods employing structural information that have been successfully applied to high-resolution models [33-36]. Others have also explored the range of applicability of low-resolution structures for functional inference [37]. Most recently, for a large set of enzymatic functions, Arakaki and coworkers demonstrated that structures whose backbone RMSD < 4 Å can be useful for biochemical function prediction with the accuracy significantly increasing for structures with a RMSD below 3 Å [38]. Therefore, the development of approaches that can refine low resolution structures to higher resolution structures at

atomic detail is highly important, not only for biochemical functional prediction, but also for ligand screening [39,40].

The problem of protein refinement has garnered much less attention so far and still poses significant difficulties. But it has already been appreciated as a limiting step towards high-resolution protein structure prediction [41,42]. This recognition resulted in the launch of the new category of CASP-Refinement (CASPR) in the last edition of the challenge. So far, there have been only a few attempts at all-atom protein folding [43-47] and protein model refinement [15,48-55]. Protein folding simulations with atomic details were conducted only for single small, fast folding proteins due to the very high computational cost. The refinement studies also report only isolated instances of structure improvement. Often the refinement attempts fail [49], or succeed only for a very small fraction of tested models [53]. The use of explicit solvent coupled with long simulation time does not help with the refinement quality [53]. Similarly, as for the coarse-grained simulations, restraints from a homology model can be used to limit the conformational search [55]. With the best improvement being about 2 Å [51], the methodology is far from routine, and to date has mainly been applied to very small protein systems. Thus, there are considerable issues that must be addressed in order to improve the state of the art.

In our research, we evaluated the performance of existing all-atom force fields for the task of protein model refinement (Chapter 2) and employed a powerful global optimization method to sculpt a funnel-shape for the best performing potential (Chapter 3). Finally, we report the results of first systematic refinement that we were able to obtain with our new, optimized force field for a representative and large set of proteins and decoys (Chapter 4).

CHAPTER 2

CAN A PHYSICS–BASED, ALL–ATOM POTENTIAL FIND A PROTEIN'S NATIVE STRUCTURE AMONG MISFOLDED STRUCTURES? I. LARGE SCALE AMBER BENCHMARKING

2.1 Introduction

With the improvement of protein structure prediction methods, the protein model refinement problem is becoming increasingly important. State of the art structure prediction procedures, including *TASSER* [12,13], *ROSETTA* [56], *PCONS* [57], *3D-SHOTGUN* [58] or *CABS* [14] are able to assemble approximately correct structures for a significant fraction of protein sequences when a weakly homologous structure is available in the Protein Databank, PDB [59]. In a benchmark test for proteins covering the PDB below 35% sequence identity, *TASSER* was able to predict models with a root mean square deviation from native, RMSD, < 6.5 Å for ~70% of single domain proteins < 200 residues in length, and ~60% proteins of < 300 residues [13,60]. Yet while these results are encouraging, the models are generally not close enough to native for use in biochemical function prediction or ligand screening as part of the drug discovery process [38]. This fact highlights the importance of developing approaches that can refine low-resolution structures to higher resolution. A natural choice for a refinement protocol would involve a detailed atomic model and the use of all-atom physics based potentials. There has been some work in the direction of both structure ranking and refinement using all atom potentials over the last decade. The *AMBER* potential [61,62] assisted by different solvation schemes was tested by Lee [49], Hsieh [63] and Lee [64] for their scoring ability to rank a set of structures. This work showed that *AMBER* could recognize the native structure among a

variety of decoys with a good accuracy. Lazaridis et al [65] and Dominy et al [66] tested the *CHARMM* [67] potential in a similar way using both decoys generated in folding experiments by other force fields, and decoys of the native fold with sequences borrowed from different proteins. In such tests, *CHARMM* also successfully scored the native structure as having the lowest energy. Similarly, the *OPLS* force field [16] was shown to have the ability to find the native structure among a set of misfolded structures [68]. Additionally, there are examples of successful native ranking with knowledge based and simplified all-atom potentials [15,69,70].

In contrast to the promising results of structure ranking of conformations generated by alternative protocols, the case of structure refinement has seen much less success. The few reported examples include the work of Vieth [51], Samudrala [71], Simmerling [48], Lee [50], and Bradley [15] with the best improvement being about 2 Å [51]. The common explanation for the discrepancy between the scoring and refinement results is that the conformational search using an all-atom force field is computationally very demanding; thus, the requisite CPU times to achieve such an improvement are excessive. Of course, underlying this statement is the belief that extant atomic potentials are adequate and the problem is merely one of the conformational search.

For any given potential to be suitable for structure prediction or refinement, there are two conditions that need to be fulfilled: 1.) the potential must score the native structure as the lowest in energy and 2.) there must be a correlation of the potential energy with native-likeness (e.g. RMSD) to drive the conformational search in the direction of the native structure. By exploring the energy surfaces of 150 single domain proteins using an all-atom (*AMBER*) potential, we try to answer whether the *AMBER*

potential fulfills the above conditions, and therefore whether it can be used for structure refinement. In contrast to previous studies, here we address the issue of how the conformational search, driven by the *AMBER* potential, affects the scoring results. The search is applied to both native and a set of decoy structures that span the range of significant to random relationships to the native structure. Also, we present results for the largest testing set of proteins used so far. In order to account for the solvation component of the free energy, we use *AMBER* with the generalized Born (GB) implicit solvation model [28] and also include a surface area dependent term (SA).

Using a representative set of 150 proteins and their associated protein-like decoys, we monitor the ranking of the native structure and investigate the relationship between native similarity and the energy of *AMBER/GBSA* potential as a function of search time. We conducted the tests in three different Relaxation Regimes: I. at time zero, with only minimization of all the native and decoy structures, II. after local relaxation: a 200 ps molecular dynamics (MD) search was conducted, followed by minimization, and III. after a relatively extensive search: 2 ns of MD, followed by minimization. The objective is to see how the extent of the conformational search affects the scoring results, and what is the shape of configurational free energy space for all the proteins in different relaxation windows. Finally, our goal is to answer the key question: is the search problem the main reason for the slow progress in the all-atom protein structure refinement field?

2.2 Methods

2.2.1 Protein sets

In this study, we employ a previously prepared [72] comprehensive benchmark protein set, which includes 1489 test proteins and covers the PDB library [59] with lengths from 41 to 200 residues at 35% sequence identity (PDB200). Both the native structure and a collection of protein-like decoys from *TASSER* are available for each protein in the set. We then randomly select a subset of 150 proteins from the PDB200 set according to following criteria: 1.) *The structures do not contain large ligands, prosthetic groups, and binding partners necessary for maintaining the fold.* Such a selection is justified by the fact that we cannot include crystallization partners in the calculation; we found that most of the structures co-crystallized with large partners were not stable when subject to molecular dynamics based relaxation. 2.) *The structures were obtained by X-ray crystallography*. We also decided not to include most NMR structures due to their conformational ambiguity. NMR structures are usually deposited in the PDB library as a collection of models that satisfy spatial restraints from experiment. Typically, the models are composed of a structurally conserved core region and variable regions (loops and chain ends). The variable regions may cause structural differences as large as 5 Å in the Cα RMSD from native, and the collection of models covers a large spectrum of *AMBER* energies. Since our goal is to compare the native energy with the energies of decoy structures, it is crucial to have one, well-defined native conformation. Therefore, for further calculations, we use only a few NMR structures, for which a structurally close (RMSD < 2 Å) X-ray mutant or homologue structure is available in the PDB.

Some of the proteins from our testing set were crystallized as part of larger molecular assemblies but we verified by comparing different PDB entries that the same protein or a close homologue had a very similar structure despite different crystallization conditions. Our assumption was that the partners are not essential for maintaining the fold in such cases.

The set of 150 proteins will be denoted as the "150-set". Since Regime III is computationally very demanding, we were not able to explore it for the whole 150-set. Thus, we selected 50 smaller proteins from the 150-set, termed the "50-set" in what follows. The PDB ID list for both sets is available at:
http://cssb.biology.gatech.edu/skolnick/files/all-atom/.

2.2.2 Decoys

The decoys used in this work come from *TASSER*-based protein structure prediction [13]. These decoys have protein-like topologies and interactions, yet they vary in their similarity to the native structure. *TASSER* uses a coarse-grained protein model of two interaction centers per amino acid (the Cα and side chain center of mass, CM). All atom structures of 14,000 decoys per protein were constructed using *PULCHRA* [19]. Then, the decoy set for a given protein was divided into 50 intervals of descending native similarity measured by the TM-score [72], and up to 20 models per each interval were chosen for further calculations (giving up to 20x50 decoys per protein). All decoy structures were minimized, and the lowest energy decoy from each TM-score interval was selected for further calculations. Separately, for a few proteins, all 14,000 decoys were minimized with *AMBER* and used for comparison. The results from both protocols are very similar, and therefore, the use of the less time consuming protocol is justified. A

side chain reconstruction procedure was also applied to the native structure: the native structure was first reduced to a Cα + CM (side chain center of mass) representation and then reconstructed with *PULCHRA* [19]. The native-reconstructed structure was also included in further calculations for comparison.

2.2.3 Structure similarity metrics

We use two different metrics to measure structure similarity: root-mean-square deviation between two structures, RMSD, and the template modeling score, TM-score [72]. While there are a variety of other structure comparison metrics that could also be used, the RMSD and TM-score metrics are chosen as they capture most of the structural similarity features we want to monitor. The RMSD is commonly used and well recognized and it appropriately describes the region of close structural similarity. But in the region of lower structural similarity, the information given by RMSD is very limited. For instance, a single hinge motion between two parts of a molecule (e.g. two domains) can lead to very high RMSD values, despite the structural similarity being otherwise very high. The RMSD is also protein size dependent. The TM-score, on the other hand, has no protein size dependence and finds the superposition of two structures that balances the coverage of the region of the protein with highest structural similarity and the alignment accuracy [72]. It weights close matches higher than distant matches. A hypothetical match of two structures that have 80% of their structures identical, but have a significantly different conformation of a terminal tail is an example. The RMSD between such structures can be very high, while the TM-score will denote a significant structural match. The TM-score ranges from (0,1], with 1 denoting identical structures. A

TM-score higher than 0.4 indicates a meaningful structural similarity and lower than 0.17 means a random match. The definition of TM-score can be found in the Appendix A.

2.2.4 Free energy function

All calculations were performed using the *AMBER* force field, ff99 [61] including GB/SA implicit solvation. In this approach, the solute (protein) is represented by an all-atom detailed model, while the solvent is treated as a mean electrical field approximated through generalized Born theory [28]. Non-polar solvation interactions are modeled by a term proportional to the solvent accessible surface area (SA) [30]. The *AMBER/GBSA* free energy is then approximated as a sum of two terms: the internal energy of the protein (the molecular mechanics energy, E_{MM}) and the solvation free energy (ΔG_{solv}), that is further decomposed into polar (ΔG_{GB}), and non-polar (ΔG_{SA}) contributions. The internal configurational entropy of the protein is neglected based on earlier predictions, that the internal entropy of a protein is similar in native, misfolded and denatured states [31]. E_{MM} is the sum of an internal strain energy (vibration of covalent bonds and rotation of valence bond angles and torsional angles), a Van der Waals energy modeled by a Lennard-Jones potential and a protein electrostatic energy approximated as Coulomb interactions of atomic point charges.

2.2.5 Scoring Protocols

Three protocols were used to establish the scoring abilities of *AMBER/GBSA* potential and to monitor the dependence of the results as a function of conformational search time. The first protocol included a short, simple minimization of the native structure (termed "native-I") and all the decoys. In the course of minimization, the structures were first relaxed, with their Cα positions frozen for 50 steps using a distance

dependent dielectric constant, to remove bad side-chain contacts that often appear in the decoy structures. Then, GB/SA solvation was turned on, and the minimization was carried out for 5,000 more steps. The TM-score span of decoys for each protein was divided into 50 intervals, and the lowest energy decoys from each TM-score interval were used for scoring analysis and further calculation. This protocol includes only local relaxation of protein structures and we refer to it further in the text, as Relaxation Regime I. Then, a more crucial test was applied: the 50 lowest energy decoys of different native structure similarity, and the native structure were subject to a 100 ps equilibration and a 100 ps molecular dynamics (MD) production run with *AMBER/GBSA*. 20 snapshots from each MD run were minimized (5,000 steps) and taken for subsequent scoring analysis; we term this Relaxation Regime II. The lowest energy snapshot from the native trajectory is further referred to as "native-II". The third protocol, as it is the most computationally demanding, was performed only on a part (50) of the 150-protein set. The 50 proteins were chosen mostly randomly, with some preference to include small proteins, and proteins with a representative (average) scoring result in Relaxation Regime II. For this subset, the MD simulation was extended to 2 ns. Again, 20 snapshots from the last part of each MD run were chosen for minimization and energy versus TM-score scoring. This protocol involves thorough relaxation of native ("native-III") and decoy structures and is denoted as Relaxation Regime III.

For all the native and decoy structures chosen for scoring, the energy gap between native structure and the lowest energy decoy is calculated to check if the force field is able to pick the native structure from a set of decoys based on energy. To test the ability of the force field to refine protein models, the correlation of the energy versus TM-score

was also monitored. A funnel like energy landscape with a good energy – native-likeness correlation would promote structural changes towards native during the conformational search. On the other hand, a flat and rugged energy landscape would trap decoy structures in local energy minima and prevent structural changes in the direction of the native state.

2.3 Results

2.3.1 Relaxation Regime I: scoring after minimization

2.3.1.1 Native – decoy energy gap (native scoring)

In the least demanding test, the energies of minimized decoy structures are compared with the energy of the native-I structure (the minimized experimental structure). We find that the energy of the native-I structure is lower than the energies of all the decoys in 100% of the cases for the considered protein set. The average native-I – decoy energy gap (the difference between native-I energy and lowest decoy energy) is -406 kcal/mole. A representative plot of energy versus TM-score for all the structures of a given protein sequence is shown in Figure 2.1A for the protein 1ag6_. Of the proteins with X-ray determined structures that we tested, the only cases in which the scoring test fails (when the native-I structure is not the lowest energy minimum) are proteins that were co-crystallized with large partners like DNA, prosthetic groups, protein ligands. These partners were not included in the energy evaluation; this is a likely reason for the failure to identify the native conformation as the lowest energy structure. Therefore, we exclude such proteins from our testing set of 150 proteins.

Interestingly, most NMR native structures were not ranked first in the set of decoys. Out of a collection of NMR models available for a given sequence in PDB, we

always evaluated only the first model and it was not always the lowest energy structure of all deposited NMR models. On this basis, we decided to exclude most NMR structures from our set of 150 testing proteins. The few NMR structures that had a structurally close homologue, determined by X-ray crystallography in the PDB, passed this test (native was ranked #1) and therefore we use them in further calculations.

The native-I – decoy energy gap was also decomposed into the contributions from different kinds of interactions. Each individual component of the *AMBER/GBSA* native-I energy is on average lower than for the lowest energy decoy structure, with the two exceptions being the GB energy term and electrostatic interactions of bonded atoms (1-4 interactions). These two components are on average lower for the lowest energy decoy than for the native-I structure. When electrostatic interactions are considered together (electrostatics + GB + electrostatics of bonded atoms), they also favor the native-I structure.

The decoy structures that come from *TASSER* have only the coordinates of the Cα atoms and side-chain centers of mass (CAS model) and the missing atoms were added before the *AMBER* energy was calculated, using *PULCHRA* [19]. We wanted to check how much the all-atom building procedure increases the energy of the native structure. We therefore converted all native structures to the CAS representation, and then applied the same reconstruction procedure as was used for decoys. Structures generated this way have on average higher energies than the original native-I structures, with an energy gap of 104.8 kcal/mole. The energy gap is significant, but much smaller than the native-I – decoy gap. These rebuilt-native structures are not considered in further calculations.

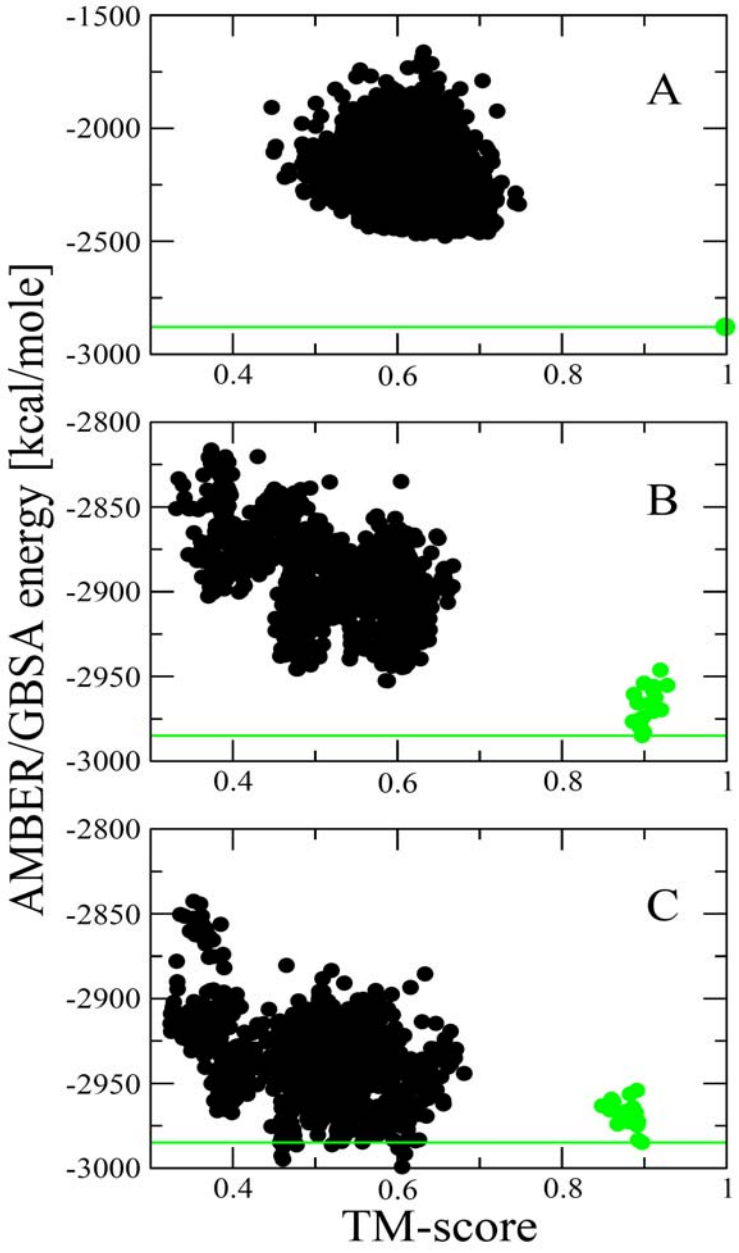


Figure 2.1 Representative plot of *AMBER/GBSA* energy as a function of TM-score in the three different Relaxation Regimes for protein 1ag6_. Green line denotes "native" energy level, black dots – decoy structures, green dots – "native structures" (minimized experimental structure or 20 minimized snapshots from MD simulation of the experimental structure). **A**: Regime I; the native structure, native-I, is the minimized experimental structure, **B**: Regime II; the native structure, native-II, is the lowest energy snapshot from 200 ps MD, starting from the experimental structure and **C**: Regime III; the native structure, native-III, is the lowest energy snapshot from 2 ns MD, starting from the experimental structure. Note the energy scale change between Regime I, and II and III.

2.3.1.2 Correlation coefficient of energy and native-likeness

In the second part of the analysis, the correlation coefficients (CC) between the TM-score (RMSD) and all energy components were monitored. The purpose was to check if *AMBER/GBSA* energy components promote native like structures among decoys. Note, we exclude the native-I structures themselves from this calculation. There is practically no correlation of energy with either TM-score or RMSD observed for decoy structures. The average correlation coefficient equals 0.4. The highest correlation is observed for the Van der Waals energy term (0.5). For about 1/3 of the proteins, the observed correlation coefficient is higher than 0.6, and in this group, the only energy term that exhibits a significant correlation with TM-score (RMSD) is again the Van der Waals term. All scoring results, also including those in Relaxation Regimes II and III, are presented in Table 2.1.

Table 2.1: Summary of the results from Relaxation Regimes I-III[a].

	Regime I		Regime II		Regime III
	150 set	50 set	150 set	50 set	50 set
% of Proteins with native energy ranked #1	100%	100%	70%	66%	20%
native – decoy energy gap	-406.05	-335.98	-25.87	-19.63	14.44
<native energy>	-4553.86	-3863.69	-4768.47	-4060.43	-4092.73
<lowest decoy energy>	-4147.81	-3527.71	-4742.60	-4040.80	-4107.17
energy - TM-score CC	0.43	0.39	0.18	0.13	0.08

[a]All energies are given in kcal/mole and the correlation coefficients (CC) are given for decoy structures only.

2.3.1.3 Decoy scoring

When the native-I structure is compared to the decoys, the native-I always has the best energy and best TM-score. Next, we ask whether the best decoy structure can be

chosen by the best energy. Only in 4 cases does the lowest energy decoy have the best TM-score. All decoy-scoring results are presented in Table 2.2.

Table 2.2: Summary of the decoy scoring results for Relaxation Regimes I, II, III.

	Regime I		Regime II		Regime III
	150 set	50 set	150 set	50 set	50 set
<TM-score> of the best decoy	0.62	0.59	0.59	0.56	0.54
# of proteins with the best decoy ranked #1	4	1	0	0	0
<TM-score> of the lowest energy decoy	0.56	0.52	0.48	0.43	0.39

2.3.2 Relaxation Regime II: scoring after 200 ps of MD

2.3.2.1 Reference structure

In the second test, we relax all the structures with a total of 200 ps of molecular dynamics (MD) with *AMBER/GBSA*, and then we repeat a similar analysis as in Relaxation Regime I. For all proteins, after 200 ps of MD there are decoys that are lower in energy than the minimized experimental structure, native-I. Also the minimized snapshots from the native MD trajectories have lower energies than the corresponding native-I structures. Clearly, the minimized experimental structure can no longer be used as a reference. We then chose 20 snapshots from each trajectory (native and decoy) in the same time frame and compare decoys to the lowest energy snapshot from the native trajectory (native-II).

2.3.2.2 Native – decoy energy gap (native scoring)

After a short relaxation with MD, all the structures are lower in energy. However, for the decoys, the energy decrease is much more pronounced than for the native-II

structures. Now, the lowest energy structure comes from the native trajectory in 70% of the cases. Thus, on average, the native-II structures are still lower in energy than the decoys, but the native-II – decoy energy gap is much smaller than in Relaxation Region I and is now –25.9 kcal/mole. A representative plot of energy versus TM-score is shown in Figure 2.1B for 1ag6_. In the course of molecular dynamics, the native trajectory deviates from the experimental structure, and in the case of some proteins, the native-II structures are of similar quality in terms of their RMSD from the experimental structure as the best decoys. That is, the structures that started from the experimental native structure begin to drift away. We then additionally apply a cutoff for "nativeness" of a 2.5 Å RMSD from the experimental structure. All proteins with the native-II structure above the RMSD cutoff and lowest energy decoys below the cutoff are discarded. This way, we ensure the reference structure is always within 2.5 Å from the experimental structure and the decoy structure does not belong to our arbitrarily chosen native cluster. Such a filtration process leaves 118 proteins for further analysis. In this set, the lowest energy structure comes from the native trajectory in 75% of the cases, and the average native-II – decoy energy gap is –26.28 kcal/mole. Whether the filter for "nativeness" is applied or not, the two analyses give similar results, which is also indicative of their robustness. Also, decomposition of the energy gap into different energy terms is consistent in the two analyses, and the energy terms that consistently favor native-II structures are the bond stretching, angle bending and Van der Waals energy terms. On the other hand, the dihedral angle energy term consistently favors decoy structures. All the other components display almost no preference towards native-II or decoy structures (they favor native-II in

nearly half of the cases), but on average, the electrostatics and SA terms favor decoys and the GB term favors native-II structures.

The energy gap correlates best with the energy gaps calculated for the bond, angle and Van der Waals components. It also correlates with the number of atoms in the protein (the larger the protein, the more negative the energy gap). There is no correlation between the total energy gap and the quality (RMSD) of the lowest energy decoy. The average RMSD of the decoys that are lower in energy than the corresponding native-II structures is 6.9 Å and 1/3 of the decoys have a RMSD higher than 10 Å. For the cases of proteins when the native-II has the lowest energy, the average RMSD of the lowest energy decoy is 8.2 Å.

2.3.2.3 Correlation Coefficient of energy with native-likeness

There is no correlation observed between the RMSD of the decoys and their energy. The average correlation coefficient, CC, equals 0.18, and only for 2 proteins is the correlation higher than 0.6. There is also very little correlation of energy with RMSD for native snapshots. The average correlation coefficient is even smaller (CC=0.03) but there are 17/150 proteins for which the correlation is higher than 0.6. The results are very similar, when the TM-score as a measure of native similarity is used instead of the RMSD.

2.3.2.4 Decoy Scoring

The best decoy structures after relaxation with MD have on average lower native similarity than the best starting decoys. Also, the best TM-score decoy is never the lowest energy decoy.

2.3.3 Relaxation Regime III: scoring after 2 ns of MD

2.3.3.1 Reference structure

In the course of the MD simulation, lower energy states are found in both the native and decoy trajectories. The average improvement in energy for the minimum energy native snapshot over the native-II structure (minimum energy native snapshot from the Relaxation Regime II) is ~30 kcal/mole. We again use the lowest energy snapshot from last 100 ps of the native 2 ns MD simulation as the reference point, and we refer to it as native-III. Only in the case of one protein, 1ag6_, was the Relaxation Regime II snapshot lower in energy by 1.5 kcal/mole, and therefore it is used as the reference native energy.

2.3.3.2 Native – decoy energy gap (native scoring)

In the most demanding of our three tests, the average native-III – decoy energy gap is no longer favorable for the native-III structures, and on average, it is +14 kcal/mole. For most proteins (80%), one of the higher RMSD (non native) decoys is the lowest energy structure. A representative result is shown in Figure 2.1C for 1ag6_. We also checked the results using a 2.5 Å RMSD cutoff for "nativeness", as in Relaxation Regime II, but this does not significantly change the results. When the energy gap is decomposed into contributions from the different energy components, we again observe that the terms that most favor the decoys are the dihedral angle energy term and the electrostatics of bonded atoms (1-4 interactions). Also, the SA energy term often favors the decoys over native-III, but its contribution is very small. Bond, angle and Van der Waals interactions of bonded atoms (1-4) consistently favor native-III structures. Also, when the energy gap is calculated separately for proteins where the lowest energy decoy

structure has a higher energy than the native-III structure and those whose decoys have energy lower than native-III, the largest difference between the two sets is in the Van der Waals interactions. This contribution is on average much more favorable towards native in those cases when the native-III structure is the lowest energy structure. The electrostatics and GB solvation terms favor native-like structures in approximately half of the cases, and they always contradict each other. On average, electrostatics favors decoy structures and GB favors native-III structures, but both contributions are large and they have large error bars (the addition of another protein to the testing set can completely change the average result).

2.3.3.3 Correlation coefficient of energy and native-likeness

The energy does not exhibit any correlation with either the RMSD or TM-score; indeed for all of the proteins, the correlation coefficient is lower than 0.5, and on average it is 0.1. A weak correlation (0.5-0.6) is observed only for bond, angle and Van der Waals energy terms in the case of a few proteins. For other components, the correlation coefficient is always lower than 0.5. A distribution of the correlation coefficients for our protein set is presented in Figure 2.2.

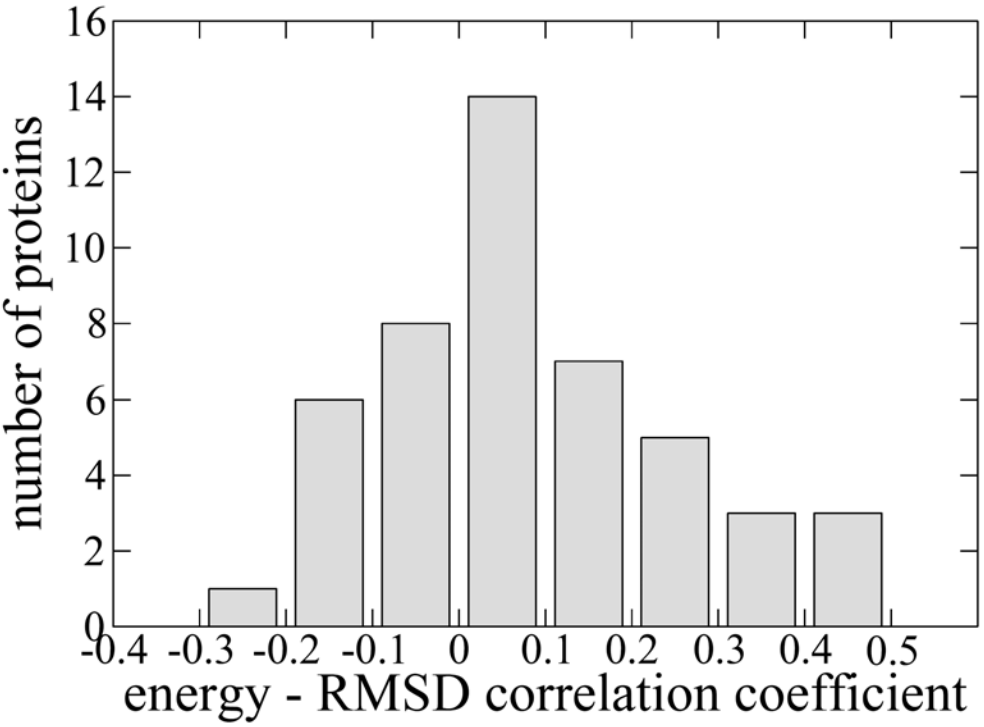


Figure 2.2 Histogram representation of the correlation coefficients between *AMBER/GBSA* energy and RMSD from the experimental structure for Relaxation Regime III.

2.3.3.4 Decoy Scoring

After 2 ns of MD, the best TM-score decoy is never the lowest energy decoy. Also, the best energy decoy has an average TM-score of 0.39, indicative of rather weak structural similarity to the native structure. Clearly, the longer the search, the lower the energy minima that are found for decoys distant from the native structure.

2.3.4 Drift of decoys

We also monitor the RMSD and TM-score change of the decoys relative to the native structure in the course of the MD simulation. Using RMSD as the structural similarity measure, this analysis shows its drawbacks. While there are many structures

that improved relative to the native structure with the largest improvement of ~ 10 Å towards native, most changes are actually meaningless and account only for a change in the decoy compactness. The RMSD improvement correlated very well with the quality of the initial decoy (the worse the decoy was initially, the larger was the change towards native). The greatest improvement was from a RMSD of 22 Å to 12 Å from native. However, the decoy structures bear no similarity to the native structure either at the beginning or the end of the MD simulation. Thus, we use the TM-score that can distinguish between meaningful and random structural changes. An improvement was found for ~15% of the decoys. The most significant improvement was of 0.1168 TM-score units. The accompanying RMSD improvement was only 0.2 Å.

2.3.5 Optimization of the *AMBER* potential

In the scoring analysis, we observed that some energy components consistently favor decoy or native structures. We next checked if, by changing the relative contributions of the individual energy components, we could improve the energy – native-likeness correlation. We use the CERN MINUIT program [73] for optimization using the optimization function as described previously [12]. In short, the function changes the weights of the energy components to maximize the correlation coefficient between energy and TM-score, and maximizes the native-decoy structure energy gap. The optimization attempt was unsuccessful, and the obtained improvement of correlation was essentially meaningless: the original average correlation coefficient of 0.1 increased to 0.3 on optimization. This shows that the lack of ability of the force field to recognize native-like structures arises from the parameterization of the potential, rather than from a wrong balance of energy components.

2.4 Discussion

The ability of the *AMBER/GBSA* potential to recognize the native structure among decoys in the first scoring test, when all structures are only minimized (Relaxation Regime I), appears to be an artifact of the decoy preparation procedure. Short relaxation of all the structures with 200 ps MD (Relaxation Regime II) also does not reveal the true shape of the potential. In fact, this is the way that the assessment of all-atom potentials were previously done [64,66,68]. Experimental structures are compact and have physical, well-minimized distances and angles. The decoy structures are not only misfolded, but they often contain unrealistic conformations of side chains and have much worse packing than experimental models. When decoys and experimental structures are only minimized prior to energy comparison, the challenge for a scoring function is mainly to recognize the most compact and best-packed structure, rather than a true native fold. This is also why the Van der Waals energy is the only contribution that correlates with native similarity in our test for Relaxation Regime I. But when all the structures are well relaxed with the scoring potential, prior to energy comparison, the differences in compactness and packing disappear, and it becomes a real challenge to select the native-like structure from a collection of alternatives. In such a test, the *AMBER/GBSA* energy fails, and the results reveal that the potential has a quite flat and rugged landscape, with many comparable minima away from the native structure. The results (average native – decoy energy gap) do not depend on the quality of the decoys used for the analysis; in the set of proteins with a decoy structure lower in energy than the native-III structure, the lowest energy decoy was ~3-16 Å away from the native structure. The results also do not depend on the secondary structural class of protein used. The average native – decoy energy gap

depends however on the length and temperature of the MD simulation (results for the temperature dependence are not shown). The more thorough the search, the lower the energy minima of both non-native and native decoys that are found. Indeed, the energy landscape seems rather flat, since no correlation is found between energy and native-likeness, even in the near native region. Previous work has also indicated some possible improvements to the accuracy of the *AMBER/GBSA* potential. These factors include optimization of the dihedral angle parameters and partial charges [26,46,74-78], correcting the generalized Born solvent model [29,79,80], and the development of improved functional forms to model nonpolar solute – solvent interactions [81].

An attempt to optimize the weights of the *AMBER/GBSA* potential shows that a simple balancing of the different energy terms does not significantly improve the scoring abilities of the potential. In order to change the potential landscape into a funnel-like shape with the minimum corresponding to the native structure, changes in the force field parameters are needed. Decomposition of native – decoy energy gaps and energy – native-likeness correlation coefficients into contributions coming from different interactions may help guide the optimization procedure. For example, the dihedral angle component consistently favors decoy structures over native. Also the electrostatic energy, that is always a large contribution, does not help to distinguish between native and decoy structures at all. An extension of such analysis is a comparison of *AMBER* (ff99) results with the results from different force fields for the same set of proteins. In the next chapter (Chapter 3) we will explore the performance of the newer, ff03 version of *AMBER* potential [74].

CHAPTER 3

DEVELOPMENT OF A PHYSICS - BASED FORCE FIELD FOR THE SCORING AND REFINEMENT OF PROTEIN MODELS

3.1 Introduction

Two of the major unsolved problems in protein structure prediction involve the scoring of decoy structures such that the most native-like conformation is selected on the basis of its energy, and the refinement of low-resolution protein models to higher accuracy [55]. In practice, the correct scoring of decoys is a less complex task than their systematic refinement [82]. Indeed, for a significant fraction of tested proteins, many potentials correctly identify the native structure as having the lowest energy among decoys [15,63-66,68-70]. However, only very rarely is there a correlation between energy and native similarity [12]. For such typical predictions, this correlation is necessary for choosing the decoy closest to the native structure on the basis of its energy and most likely represents the physically realistic situation.

The refinement of low-resolution predicted models with a backbone root mean square deviation from the native structure, RMSD, of about 6 Å, to high-resolution all-atom structures whose RMSD is less than 2 Å has proven to be an extremely difficult task. The solution to this problem has become more essential with the improvement of protein structure prediction methods. State of the art structure prediction procedures, including *TASSER* [12,13], *ROSETTA* [56], *PCONS* [57], *3D-SHOTGUN* [58] or *CABS* [14] generate approximately correct structures for a significant fraction of protein sequences for which a weakly homologous structure is available in the Protein Databank, PDB [59]. For example, in a benchmark test for proteins covering the PDB below 35% sequence

identity, *TASSER* was able to predict models with a RMSD < 6.5 Å for ~70% of single domain proteins < 200 residues in length, and ~60% proteins < 300 residues. However, for many important applications such as detailed studies of interactions, molecular mechanisms, ligand screening and drug design, more accurate structures at atomic detail are required.

For structure refinement to be routine, the correlation of energy with native-likeness has to be satisfied not only for the ranking of decoys generated in an extrinsic procedure by a different energy function but also for the collection of structures generated when the energy function drives the search. As we demonstrated [82] in Chapter 2, the apparent correlation of energy vs. native-likeness observed for one potential when the decoys are generated with another potential is often an artifact of decoy preparation. Native structures are compact with well-minimized distances and angles. Decoy structures are not only misfolded, but often contain unrealistic side chain conformations with much worse packing than experimental structures. When decoys and native structures are only minimized prior to energy comparison, the challenge for a scoring function is mainly to recognize the most compact and best-packed structure, rather than the native fold. When all the structures are well relaxed with the scoring potential, prior to energy comparison, the differences in compactness and packing disappear, and it becomes a significant challenge to select the native-like structure from the sea of alternatives [82]. For a set of 150 proteins, we have shown using the Amber ff99 potential [61,62] and decoys obtained with the *TASSER* force field [12,13], that a weak correlation (~ 0.4 on average) of the energy with TM-score, (a measure of structural similarity that ranges from 0 to 1.0 for identical structures, with a value of 0.3 for the best structural alignment

of a pair of randomly related structures) is observed only for the initial set of decoys [82]. Using the initial set of decoys as starting structures, after running a molecular dynamics search with the ff99 potential, this correlation decreases during the course of the simulation and is lost completely after a longer search, revealing the inherent flatness of the sampled potential. Similarly, the ability of the ff99 potential to rank the native structure as the lowest energy among initial decoys for 100% of tested proteins drops to 20% after a longer conformational search.

Among the reasons that the native structure does not correspond to the global minimum of energy for many force fields and that the correlation between the energy and native similarity is low, is that during the creation of the force field, not enough information about the global shape of the energy landscape is taken into account. Such energy global landscape sculpting was employed by Zhang at al. [12] for a large set of decoys and proteins to optimize the weights of the *TASSER* force field which employs a reduced protein model. For both sets of nonhomologous training and testing proteins, the average correlation coefficient of energy and RMSD was 0.69. A similar idea was also employed by Liwo et al. [83,84] on a much smaller set of proteins to optimize the parameters of the coarse-grained *UNRES* potential for *ab-initio* protein structure prediction. These ideas were also employed to derive an all-atom force field (*ECEPP-5*) for the prediction of the crystal structures of organic molecules [85-87].

Here, we explore the ability of global parameter optimization to sculpt a funnel-like landscape for the all-atom physics-based Amber ff03 [74] potential. For 58 nonhomologous proteins, we use a large number of decoys generated with the ff03 force field and optimize the relative weights of the energy components. We obtain a significant

improvement in the correlation of the energy with native-likeness of the decoys and the ranking of the native structure as the lowest energy as compared to the original ff03 potential [74]. Next, we show that by adding an explicit backbone hydrogen bond potential (HB) to the ff03 force field followed by global optimization of the combined potential, there is a further significant improvement in the funnel-like character of the energy landscape. We also investigated the relative contributions to the ff03 force field (supplemented by the HB potential), by turning off the electrostatic energy and generalized Born solvation [28] energy components. The optimized reduced force field still scores the native structures better than the original ff03 potential and retains the improved correlation of the energy with native-likeness.

3.2 Methods

3.2.1 Benchmarking of the ff03 force field

In this study, we use the same benchmarking protocol as described in detail in Chapter 2 for the evaluation of the *AMBER* ff99 force field [82]. In short, we employ a previously prepared [13] comprehensive benchmark protein set, PDB200, which includes 1489 test proteins and covers the PDB library [59] with lengths from 41 to 200 residues at 35% sequence identity and randomly select 58 proteins that satisfy the following criteria: 1.) The structures do not contain large ligands, prosthetic groups, and binding partners necessary for maintaining the fold, and 2.) The structures were obtained by X-ray crystallography. For these 58 proteins (listed in Table B.1, Appendix B), we take into consideration both the native and decoy structures of varying native similarity. The initial set of decoy structures were generated by *TASSER* [12,13]. 50 decoys per protein were chosen that they span the range of native similarity from essentially random to native-like

structures. All-atom representations of the decoys were constructed using PULCHRA [19]. For the native structure and all-atom decoys, we examined the performance of the ff03 potential [74] that includes generalized Born and surface area dependent solvation, GB/SA, terms [28,30] in three relaxation regimes: I. after minimization with *AMBER*, II. after 200 ps of molecular dynamics (MD) simulation, and III. after 2 ns of MD.

3.2.2 Decoys for force field optimization

To further improve the coverage of conformational space by decoys, we picked 50 low-energy decoys from MD trajectories and used them as starting structures in a thorough conformational search using the *A-TASSER* program, which is described below in the section "Conformational search method". Finally, about 30,000 decoys per protein were collected, minimized in ff03/GB/SA potential and used in force field optimization. We call this decoy set, Set58. In preparation of Set58, we required on average a low correlation of the decoys' TM-score [72] to the native state with their radius of gyration, to avoid the situation where the correlation is associated only with bad packing ("swollen" decoys). The average correlation coefficient of the TM-score to the native structure with the radius of gyration for Set58 was 0.40 (thus most decoys are well packed, compact but not necessarily native structures). There are six proteins (1ame_, 1em9A, 1a0b_, 1a7xA, 1bm8_, 1a19A) in Set58 for which the correlation of the TM-score with radius of gyration was high. We did not exclude them to increase the diversity of the decoy set and to have some representation of less well packed, "swollen" decoys in the set of structures used for parameter optimization.

3.2.3 Conformational search method

To search the conformational space of proteins and generate more decoy structures, we used our newly developed *A-TASSER* program. *A-TASSER* (for *atomic-TASSER*) represents the protein at atomic detail and employs the Replica Exchange [88,89] Monte Carlo (REMC) search method with a Parallel Hyperbolic Sampling (PHS) acceptance criterion [90] to reduce higher energy barriers. *A-TASSER* employs three types of moves that only change the torsional angles of the molecule: local "fixed end" moves [91], end moves, and the side chain moves (Figure C.1, Appendix C). The rotation angle is randomly chosen within a given amplitude range. We used (-30, 30) and (0, 360) degree rotation amplitude ranges for the end moves and the side chain torsional moves, respectively. The "fixed end" moves rotate a fragment comprised of a few residues (from 2 to 12 residues) around the axis connecting the C_α atoms of the residues at the fragment ends, whereas the rest of the protein remains unchanged. For each local move, the rotation amplitude is adjusted so that the backbone valence angles of the end residues of the fragment do not change beyond the statistical fluctuation range, which is about five degrees [91]. The amplitude of this motion typically does not exceed 30 degrees. The end moves rotate the free ends of the molecules and involve 1-5 residues. The side chain moves rotate the side chain atoms by perturbing one or two randomly chosen torsional angles. The move types and the torsional angles to be perturbed are also randomly chosen at each step. The bond lengths and valence angles do not change during the search (except for the backbone valence angles of the end residues of the fragment undergoing the "fixed end" move that are allowed to change within the statistical fluctuations seen in native proteins).

3.2.4 Force field optimization method

For each tested potential (described below in the section titled "Types of the optimized force fields"), the energy components, E_i were multiplied by individual weights, w_i, Eq.3.1, and the weights were optimized to minimize the target function F, Eq.3.2-6.

$$E_{TOT} = \sum_i w_i E_i \qquad \text{Eq.3.1}$$

$$F = G_1 \cdot G_2 \cdot G_3 \qquad \text{Eq.3.2}$$

$$G_1 = 1 - \exp(A_1 \cdot (CC - 1)) \qquad \text{Eq.3.3}$$

$$G_2 = 1 - \frac{1}{1 + A_2 \cdot \chi^2} \qquad \text{Eq.3.4}$$

$$G_3 = \frac{1}{1 + \exp(A_3 \cdot Zscore)} \qquad \text{Eq.3.5}$$

$$Z - score = \frac{\langle E_{Dec} \rangle - \langle E_{Nat} \rangle}{\left(\langle E^2{}_{Dec} \rangle - \langle E_{Dec} \rangle^2 \right)^{1/2}} \qquad \text{Eq.3.6}$$

During minimization of the function F, the component G_1, Eq.3.3, tends to maximize the linear correlation coefficient, CC, of the total energy, E_{TOT}, with the TM-score [72]. We maximized the CC only for decoys with a TM-score to the native state in the range 1 to 0.4 (structures with higher TM-score are closer to the native state). Structures with a TM-score below 0.4 are usually far from the native state, and there is no reason to expect a correlation of energy with native-likeness in this regime. The energies of these

structures are only expected to be higher than the energies of the structures closer to the native state. However, during optimization such a requirement was not explicitly enforced.

The component G_2 minimizes the deviation of the dependence of the energy on TM-score from linearity, through minimization of the chi-square value (χ^2). G_3 maximizes the gap between the ensemble of native-like structures (those whose TM-score to the native structure is larger than 0.9) and non-native structures, as a function of the Z-score, Eq.3.6. The correlation coefficient (CC), χ^2, and the Z-score in the function F are averaged over all proteins in the training set (described in the next section) and they depend on the weights w_i. The constants A_1, A_2, A_3 were set to 2, 0.01, and 0.5, respectively. The values of A_1, A_2 and A_3 were chosen so that the G_1, G_2, G_3 all change over the same range, from 0 to 1 (or close to 1 in case of G_1) and have a large gradient for the important ranges of the CC, Z-score, and χ^2. The behavior of G_1, G_2, and G_3 is illustrated in Figure C.2 (Appendix C). F possesses multiple minima in parameter (w_i) space. Therefore, we used a global optimization method [92] to find the global minimum of F with respect to the weights w_i. The method is independent of the starting values of the weights, and finds within a given range, multiple sets of weights that minimize function F. For each of 30 training subsets (described in the next section, "Training and testing protein sets"), we ran 10 independent optimization runs and collected the 5 lowest minima from all runs per subset. This way, we obtained 150 (30 x 5) sets of weights for every optimized potential. All 150 sets of weights were tested on the appropriate testing protein set.

3.2.5 Training and testing protein sets

From the set of 58 proteins (Set58), two different sets of 15 proteins were chosen randomly as training sets (Train1, Train2, Table A.1). The remaining 43 proteins with respect to each of two training sets constitute the testing set (Test1, Test2). To increase the diversity of the training set, we generated 15 subsets for Train1 and Train2, by the leave-one-out method. Thus, there are 30 (15 x 2) training subsets that were independently used for force field optimization.

3.2.6 Types of the optimized force fields

Three types of the potential energy functions were used to optimize the weights: a) the full version of the ff03 Amber potential, supplemented by GB/SA solvation, Eq.3.7,

$$E_{FF03} = w_{DIH}E_{DIH} + w_{VDW}E_{VDW} + w_{VDW1-4}E_{VDW1-4} + w_{ELE}E_{ELE} + w_{ELE1-4}E_{ELE1-4} + w_{GB}E_{GB} + w_{SA}E_{SA}$$

Eq.3.7

b) the ff03 potential with an explicit hydrogen bond potential added (HB), Eq.3.8,

$$E_{FF03/HB} = w_{DIH}E_{DIH} + w_{VDW}E_{VDW} + w_{VDW1-4}E_{VDW1-4} + w_{ELE}E_{ELE} + w_{ELE1-4}E_{ELE1-4} + w_{GB}E_{GB} + w_{SA}E_{SA} + w_{HB}E_{HB}$$

Eq.3.8

and c) the ff03 potential with HB but without electrostatic interactions and GB (omitted by setting the weights for those energy components to zero), Eq.3.9

$$E_{FF03/HB/R} = w_{DIH}E_{DIH} + w_{VDW}E_{VDW} + w_{VDW1-4}E_{VDW1-4} + w_{SA}E_{SA} + w_{HB}E_{HB}$$

Eq.3.9

Since the sampling method keeps the bonds and valence angles unchanged, we also set the weights in front of the bond and angle energy components to 0.

In equations 3.7-9, the following abbreviations are used: DIH – dihedral component, VDW – van der Waals energy, VDW1-4 – van der Waals term for atom pairs separated by less than four bonds, ELE – electrostatics, ELE1-4 – electrostatics for atoms separated by less than four bonds, GB – generalized Born energy (electrostatic component of solvation, we used the GB parameter set from Onufriev at al. [29]), and SA – surface area dependent term (hydrophobic component of solvation).

3.2.7 The hydrogen bond potential

We have tested two different approaches for the calculation of the hydrogen bond energy: 1.) a knowledge-based *TASSER*-like [12,93] hydrogen bond potential, and 2.) the *DSSP* potential [94]. Although the performance in terms of native scoring and energy-native-likeness correlation of the two potentials is very similar, the *DSSP* energy is less computationally expensive. Therefore, the hydrogen bond potential that we employ in this study follows the *DSSP* approach. The hydrogen bond energy of the system C-O · · · H-N is calculated according to Eq.3.10:

$$E_{HB} = q_1 q_2 \left(\frac{1}{r(NO)} + \frac{1}{r(CH)} - \frac{1}{r(OH)} - \frac{1}{r(CN)} \right) \cdot 332 \qquad \text{Eq.3.10}$$

where $q_1 = 0.42e$ and $q_2 = 0.20e$, with e being the magnitude of the charge on an electron, r(AB) is the distance between atoms A and B in Å, and E_{HB} is the energy in kcal/mole. A hydrogen bond occurs when two cutoff criteria are satisfied: 1.) the N-O distance is ≤ 5.2 Å, and 2.) the calculated energy is less than –0.5 kcal/mol. Only energies for backbone hydrogen bonds were calculated.

3.3 Results and Discussion

3.3.1 Comparison of scoring performance of the ff03 and ff99 potentials

Similar to our previous study [82], we performed tests of the ff03 force field in three relaxation regimes: I. after minimization with *AMBER* ff03/GBSA, II. after 200 ps of MD (followed by minimization of MD snapshot structures), and III. after 2 ns of MD (followed by minimization of the snapshots). As in the case of the ff99 force field, we found that the initial structures, the native and decoys, are in very shallow energy minima. During the conformational search with MD, much deeper minima are found nearby, and the true shape of the potential is only revealed after a long relaxation time. The most important conclusion from this initial analysis is that the ff03 force field performs better than the ff99 potential in terms of scoring the native structure as the lowest in energy and correlation between energy and native similarity. The correlation coefficient for the ff99 force field was only 0.1, while for ff03, it is 0.25. For the ff99 potential, native-like structures are the lowest energy among the decoys for only 20% of tested proteins. In the case of ff03, this is true for 48% of proteins, when a similar criterion for "native-likeness" is used (RMSD of 2 Å or less from the experimental structure). Such results are encouraging for the purpose of force field optimization, and we decided to use the ff03 potential as our base energy function in all further calculations. The optimization of the ff03 force field is required because during the MD simulations using this potential, 84% of the decoys drifted farther away from the native structure, and only 16% of the decoys improved their TM-score to the native state.

3.3.2 Correlation of energy with native-likeness in the original ff03 force field

For Set58, we calculated the linear correlation coefficients (CC) of the total energy and each energy component of the original ff03 force field, Eq.3.7, with the TM-score to the native structure. The results are shown in Table 3.1 (the hydrogen bond energy, HB is not present in the original ff03 force field). The correlation coefficient of the total energy (ETOT) is low, 0.25. Among all the energy components, the bond (BOND) and van der Waals (VDW) energy have a weak correlation with TM-score, with a CC above 0.4, whereas the remaining energy components have no correlation with native-likeness. Therefore, during optimization, one would expect the weights of these two components to dominate. Since our conformational search method fixes the bond lengths and valence angles, the bond and angle energy is set to zero during optimization. It is very interesting to note that the electrostatic interactions (ELE, ELE1-4) and generalized Born solvation energy (GB) are completely uncorrelated with native-likeness (their correlation coefficients with TM-score are close to 0). These interactions appear to be non-specific in recognizing similarity to the native structure. Therefore, one could expect relatively small values of the weights at those energy components during force field optimization.

Table 3.1 The average correlation coefficients (CC) and their standard deviations (in parentheses) of the individual components of the original *AMBER* ff03 potential with TM-score (rows ETOT - SA), and the average correlation coefficient of the DSSP hydrogen bond potential (HB) with TM-score for representative protein and decoy set (Set58).

Energy component	CC
ETOT[*]	0.25 (0.25)
BOND[†]	0.41 (0.23)
ANG[‡]	0.26 (0.33)
DIH[§]	-0.22 (0.29)
VDW[¶]	0.52 (0.25)
VDW1-4[‖]	-0.25 (0.23)
ELE[**]	0.06 (0.30)
ELE1-4[††]	0.05 (0.15)
GB[‡‡]	-0.09 (0.30)
SA[§§]	0.36 (0.26)
HB[¶¶]	0.58 (0.18)

[*] ETOT - total potential energy (*AMBER*, ff03+GBSA), [†]BOND – bond energy, [‡] ANG – angle energy, [§] DIH – dihedral angle energy, [¶] VDW – van der Waals energy, [‖] VDW1-4 – short distance van der Waals energy (for atom pairs separated by less than four bonds), [**] ELE – electrostatic energy, [††] ELE1-4 – short distance electrostatic energy (for atom pairs separated by less than four bonds), [‡‡] GB - generalized Born solvation energy, [§§] SA – surface area dependent solvation energy, [¶¶] HB – DSSP hydrogen bond energy (not present in the original ff03 force field).

There is no reason for the electrostatic energy to change monotonically with native similarity, and the native state does not have to have lower electrostatic energy than the decoys; it will be strongly protein-dependent. For our decoy Set58, on average we do not observe any correlation of the electrostatic energy with native-likeness at any range of TM-score to the native state (the average correlation coefficients in all ranges of TM-score are close to zero). There are only two examples of proteins with significant correlation (CC > 0.6) or anti-correlation (CC < -0.6) of the electrostatic energy with TM-score. In force fields, the "frozen" point charge approximation and the absence of polarization additionally introduce abnormally large fluctuations of the electrostatic energy, even for small changes of local geometry. In nature, the changes of electron density are smoother, because large unfavorable electrostatic interactions in some conformations are quenched by the polarization of electron density as well as screening by counterions. The GB solvation energy also has an electrostatic character and suffers from the same large nonphysical fluctuations as the electrostatic energy, possibly caused by the point charge approximation. The solvation energy is usually favorable for extended structures, and for some proteins, it may be weakly anti-correlated with native similarity, as the structures become more compact and less solvated. For Set58, the average correlation of GB energy with TM-score is close to zero at each range of TM-score to the native state, and it is negative and insignificant for most proteins. Only seven proteins have some noticeable correlation of GB energy with TM-score, among which five show a weak anti-correlation (CC < -0.4).

The dihedral energy (DIH) and short-distance van der Waals interactions on average appear to be weakly anti-correlated with native-likeness; however, their CC

values are practically negligible. The dihedral energy landscape is flat for a wide range of the native similarity. Only for the near-native region (RMSD < 2 Å) is there a noticeable anti-correlation of the dihedral energy with TM-score. This result is in accordance with our earlier observation [26] that the ff03 force field has a tendency to distort the dihedral angles from their gas phase values obtained using quantum mechanical calculations for short helices and strands of polypeptides. This suggests that the dihedral energy term might require parameter reoptimization.

3.3.3 Optimized ff03 force field

We applied the optimization procedure, described in the section "Force field optimization method" to optimize the weights of the energy components of the ff03 force field, E_{FF03}, Eq.3.8. We used training protein decoy sets described in the section "Testing and training protein sets". The weights of the bond and angle energy components were set to 0, and the remaining weights were optimized without restraints. The results for the best set of weights (Wgt-0) are presented in Table 3.2. The optimized force field (column ff03 optimized Wgt-0) has a much higher average correlation coefficient (CC_{ave}) between the energy and TM-score compared to the original potential (column ff03). On average, over entire Set58 (column Set58), the CC increased from 0.25 to 0.62 for the original ff03 and optimized ff03 force fields, respectively. The values of the correlation coefficients of the energy with TM-score for each protein, for the original and optimized ff03 force fields are given in Table B.1 (Appendix B).

Besides the CC value, we also analyzed the values of the average Z-score ($Z\text{-score}_{ave}$, Eq.3.6), the fraction of proteins with a CC larger than 0.60 (CC_{fr}) (we considered

the CC ≥ 0.60 to be a significant correlation), the fraction of proteins for which the lowest energy decoy has a TM-score to the native structure higher than 0.90, $TM_{fr,}$ and the fraction of proteins for which the lowest energy decoy has a RMSD over C_α atoms to the native structure lower than 2.0 Å, $RMSD_{fr}$.

Table 3.2 Comparison of scoring performance of the unoptimized (ff03, ff03/HB) and optimized (ff03 optimized, ff03/HB optimized) force fields.

	ff03[*]	**ff03/HB**[†]	**ff03 optimized**[‡] **Wgt-0**			**ff03/HB optimized**[§] **Wgt-1**		
	Set58[¶]	**Set58**[¶]	**Train**[‖]	**Test**[**]	**Set58**[¶]	**Train**[‖]	**Test**[**]	**Set58**[¶]
CC_{ave}[††]	0.25	0.31	0.63	0.61	0.62	0.67	0.64	0.65
Z-$score_{ave}$[‡‡]	0.16	0.23	2.65	2.18	2.30	2.59	2.19	2.29
CC_{fr}[§§]	0.12	0.14	0.47	0.49	0.48	0.60	0.65	0.64
TM_{fr}[¶¶]	0.22	0.26	0.93	0.84	0.86	1.00	0.86	0.90
$RMSD_{fr}$[‖‖]	0.48	0.55	0.93	0.88	0.89	1.00	0.88	0.91

[*] Original unoptimized ff03 potential, [†] unoptimized ff03/HB potential (ff03 supplemented by hydrogen bond potential), [‡] optimized ff03 potential, weight set Wgt-0, [§] optimized ff03/HB potential (ff03 with added hydrogen bond potential), weight set Wgt-1, [¶] Set58 - the entire set of 58 proteins, [‖] Train - training protein set, [**] Test - testing protein set, [††] CC_{ave} - average correlation coefficient of the energy with TM-score, [‡‡] Z-$score_{ave}$ - average Z-score between native cluster and the remaining decoys (native cluster is defined by TM-score ≥ 0.9), [§§] CC_{fr} - fraction of proteins with correlation coefficient of energy with TM-score greater than 0.6, [¶¶] TM_{fr} - fraction of proteins for which the lowest energy structure had the TM-score to the native state greater than 0.90, [‖‖] $RMSD_{fr}$ - fraction of proteins for which the lowest energy structure had the RMSD from the native state less than 2 Å.

The more positive the Z-score, the better is the energy separation between the native and non-native decoys clusters. The force field optimization improved the average Z-score from 0.16 to 2.30, for the original and optimized ff03 force fields, respectively. The fraction of proteins with a significant correlation coefficient, CC_{fr}, also greatly increased from 0.12 to 0.48 for the original and optimized ff03 force fields, respectively. This means that for about 48% of the proteins, selecting the lowest energy decoys guarantees that the decoys are closest to the native structure. TM_{fr} and $RMSD_{fr}$ describe the ability of a force field to pick the native structure among decoys using an energy criterion (TM-score greater than 0.90), and to indicate by energy the near-native cluster (RMSD less than 2.0 Å). The TM-score, unlike RMSD, is chain length independent, so that the two measures cannot be directly compared, but for our set of proteins and decoys a TM-score of 0.9, roughly corresponds to an average RMSD of 1.4 Å. For set 58 the TM_{fr} value increased after optimization of the force field from 0.22 to 0.86, and the $RMSD_{fr}$ increased from 0.48 to 0.89. It is important to notice that the potential optimized on the training protein set (**Train**, Table 3.2) is well transferable to the testing set (**Test**, Table 3.2).

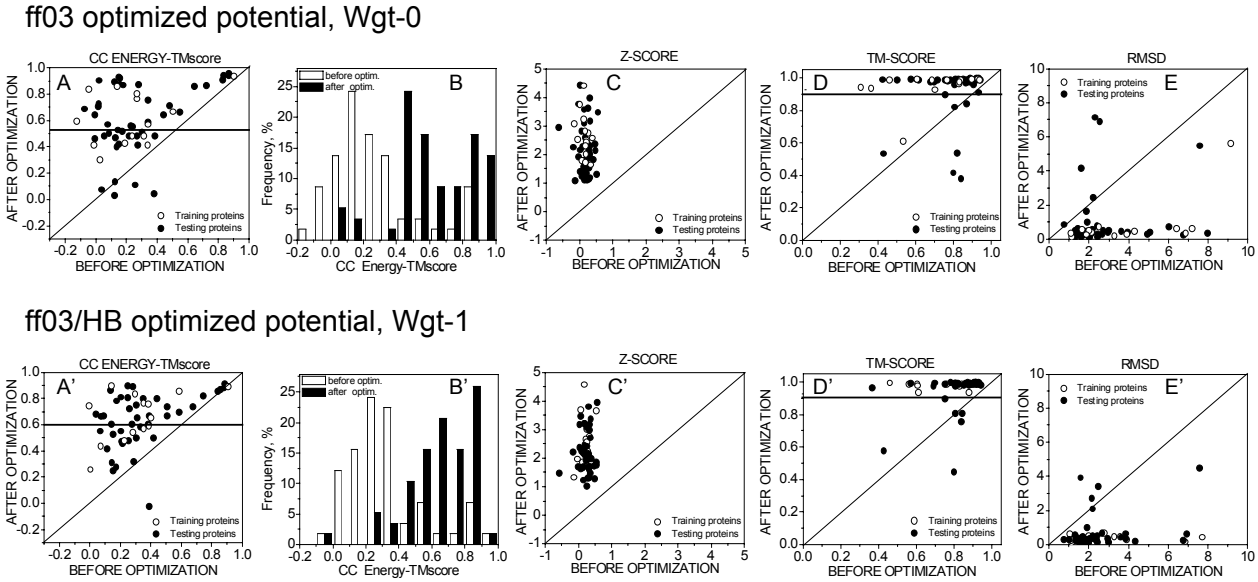


Figure 3.1 Comparison of the performance of the optimized ff03 (weight set Wgt-0) and ff03/HB (ff03 with added hydrogen bond potential, weight set Wgt-1) force fields for the set of 58 proteins (Set58), A-E – the results for the optimized ff03 potential, A'-E' – the results for the optimized ff03/HB potential, A, A' – correlation coefficients of the energy with Cα atom TM-score to the native structure after optimization with respect to the values before optimization, B, B' – distribution of correlation coefficients of the energy with Cα atom TM-score to the native structure before (open bars) and after (black bars) optimization of the force fields, C, C' – Z-score after optimization with respect to the values before optimization, D, D' – Cα atom TM-score to the native state of the lowest energy decoy after optimization with respect to the values before optimization, E, E' - Cα atom RMSD to the native state of the lowest energy decoy after optimization with respect to the values before optimization, open circles – results for the training protein set, black circles – results for the testing protein set.

Additional illustration of the performance of the optimized ff03 potential with respect to the original one is given in Figures 3.1A-E. Figure 3.1A presents the values of the correlation coefficient of the energy versus TM-score for each protein after optimization of the force field compared to the values before optimization. The CC

values improved for most of the proteins (points above the diagonal) for both the training (open circles) and testing (black circles) sets. The improvement of the correlation coefficient is also shown in Figure 3.1B, for different intervals of the CC values, where the bars represent the percent of the proteins with correlation coefficient in a given interval. The black bars represent the distribution after optimization of the force field, and the open bars represent the distribution before the optimization. There is a visible shift of the distributions toward the significant range of the CC values. Figures 3.1C-E show the values of the Z-score, TM-score of the lowest energy structure, and the RMSD of the lowest energy structure, respectively, for each protein after optimization of the force field with respect to the values before optimization. The Z-score values increased for all proteins (Figure 3.1C). The TM-score and the RMSD to the native structure of the lowest energy decoy improved for the majority of the proteins (Figure 3.1D, points above the diagonal for TM-score; Figure 3.1E, points below the diagonal for RMSD).

3.3.4 Influence of explicit hydrogen bond potential on the correlation of the energy with native-likeness and the scoring of the native structure

When the explicit hydrogen bond potential (HB), Eq.3.10, is added to the original ff03 force field (with weight equal 1), the performance of the force field improves. In Table 3.2, columns ff03 and ff03/HB compare the values of the correlation coefficient, the Z-score, CC_{fr}, TM_{fr}, and $RMSD_{fr}$ for the original ff03 and for the ff03 with the hydrogen bond potential included (nonoptimized). All the control values improve after adding the HB potential. However, the average correlation coefficient of the total energy with TM-score (CC_{ave}), increases from 0.25 to only 0.31, whereas the correlation coefficient of the hydrogen bond energy alone with TM-score (Table 3.1, HB) is much

larger, 0.58. Therefore, optimization of ff03/HB should allow for further improvement in the accuracy of the force field.

3.3.5 Optimized ff03/HB force field

Optimization greatly improves the accuracy of the combined ff03/HB force field. The values of the correlation coefficients of the energy with TM-score for each protein, for the unoptimized and optimized ff03/HB force fields are given in Table B.1 (Appendix B). The optimized ff03/HB (called Wgt-1) force field also outperforms the optimized ff03 potential (see Table 3.2). The average correlation coefficient (CC_{ave}) for the optimized ff03/HB Wgt-1 potential is higher than for the ff03 optimized potential (0.65 compared to 0.62, Table 3.2, column Set58). The fraction of proteins with a significant CC increased from 0.48 to 0.64, and the recognition of the native structure (TM_{fr}) and native cluster ($RMSD_{fr}$) is also better: 0.90 compared to 0.86 and 0.91 compared to 0.89 respectively. Figures 3.1A'-E' show a graphic representation of the performance of the optimized ff03/HB Wgt-1 force field. Figure 3.1A' presents the values of correlation coefficient of the energy versus TM-score for each protein after optimization of the force field with respect to the values before optimization. The correlation coefficient improved for almost all the proteins, and the improvement is on average larger than for the optimized ff03 force field. Also, the distribution of the CC has moved toward larger values, significantly more than for the optimized ff03 force field (compare Figure 3.1B' with Figure 3.1B). The Z-score improved for all the proteins (Figure 3.1C') and the TM-score (Figure 3.1D') and RMSD (Figure 3.1E') to the native state of the lowest energy structure improved for the great majority of the proteins. As an additional illustration, in Figures 3.2A-D we show examples of the plots of energy versus TM-score for the

original ff03 (unoptimized) potential and optimized ff03/HB, Wgt-1 potential. Figures 3.2A-C illustrate the average improvement of the correlation coefficient, and Figure 3.2D shows an example of a very large improvement of the correlation coefficient**.**

These results show the importance of an accurate hydrogen bond scheme for improving the correlation of the energy with native-likeness of protein decoys. Hydrogen bonding was previously shown to be a necessary requirement for the generation of protein like protein structures [95]. The hydrogen bond potential that contains an implicit angular dependence of the hydrogen bond energy is sensitive to small changes of the angular orientation of the atoms that form a hydrogen bond. This is reflected in continuous increase of the energy of the structures as their hydrogen bonding deviates from the perfect pattern and the good correlation of hydrogen bond energy with native-likeness, even in the region close to the native structure. Many well packed, but misfolded structures with distorted hydrogen bonding become higher in energy. Such a potential can help to recognize misfolded structures among well-packed decoys that are sometimes difficult to distinguish by van der Waals energy alone.

As in the case of the optimized pure ff03 potential, the optimized ff03/HB shows good transferability between the training (**Train**) and testing (**Test**) protein sets (see Table 3.2, ff03/HB optimized Wgt-1).

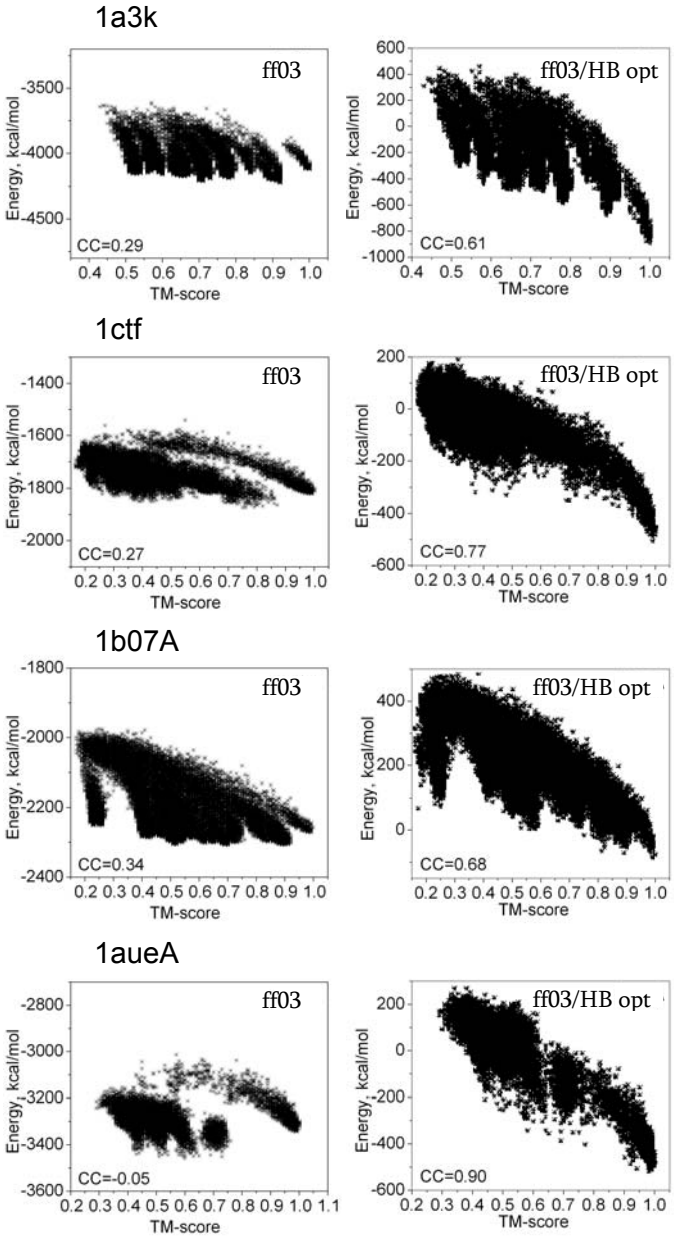


Figure 3.2 Scatter plots of the energy versus TM-score for decoy structures for the original unoptimized ff03 force field (ff03, weight set Wgt-0) and optimized ff03/HB potential (ff03/HB opt, weight set Wgt-1), A-C represent average changes of the correlation coefficient of energy with TM-score (CC), D represents a large change of the CC.

3.3.6 Weights for optimized ff03/HB force field

Among many sets of weights obtained during the optimization procedure that minimize the target function F, Eq.3.2, the best performance in decoy scoring showed the sets with some of the weights being negative for both the ff03 (Table 3.3, Wgt-0) and ff03/HB (Table 3.3, Wgt-1) force fields. The performance of these weight sets was discussed above. In the best weight set for the ff03/HB potential (Table 3.3, Wgt-1), the van der Waals, short-distance van der Waals, and hydrogen bond (HB) energies have positive and relatively large weights. The remaining weights, of the dihedral (DIH), electrostatic (ELE, ELE1-4), generalized Born solvation (GB), and surface area (SA) energy terms have negative signs. The occurrence of the negative weights for these terms indicates that they are not individually useful in generating a funnel-like shape of the potential. By assigning negative, nonphysical weights, the optimization procedure creates a linear combination of the energy terms that has larger correlation with native-likeness than the individual components. Although there is no reason for any energy component alone to have a correlation with native-likeness, and while this correlation is expected for the total energy, analysis of such individual correlations can help us to interpret the meaning of the weights in the optimized potential. In the case of the dihedral energy, the negative weight most likely reflects its initial anti-correlation with TM-score (Table 3.1, DIH). As discussed earlier (see section "Correlation of energy with native-likeness in the original ff03 force field"), the weak anti-correlation of the dihedral energy relects distortion of the backbone torsional angles in the ff03 force field from their equilibrium values, especially for near-native and native conformations. This effect was also noticed earlier by comparison with high level quantum mechanical calculations for short

polypeptide helices and strands [26]. These results suggest that the dihedral parameters may need reoptimization to better describe near-native and native conformations.

Table 3.3 Relative weights[*] of energy components for the optimized force fields.

	ff03 optimized[†]	**ff03/HB optimized**[‡]			**ff03/HB reduced optimized**[§]
	Wgt-0[¶]	**Wgt-1**[¶]	**Wgt-2**[‖]	**Wgt-3**[**]	**Wgt-R**[‖]
DIH[§§]	-1.25	-1.17	-0.32	0.28	-0.42
VDW[¶¶]	1.00[*]	1.00[*]	1.00[*]	1.00[*]	1.00[*]
VDW1-4[‖‖]	1.04	0.88	0.56	0.56	4.33
ELE[***]	-0.27	-0.40	-0.25	0.03	0
ELE1-4[†††]	-0.16	-0.23	-0.22	0.17	0
GB[‡‡‡]	-0.22	-0.23	-0.14	0.18	0
SA[§§§]	-0.51	-2.07	0.14	3.39	0.51
HB[¶¶¶]	0	6.25	1.32	2.56	4.26

[*]All weights were scaled so that the weight for van der Waals energy is equal to 1, for easier comparison, weights for bond and angle energy terms were set to 0, and are not presented; [†] optimized original ff03 force field, [‡] optimized ff03/HB force field (ff03 with added hydrogen bond potential), [§] optimized reduced ff03/HB force field (ff03 with added hydrogen bond potential, and with electrostatic (ELE and ELE1-4) and GB solvation energy components turned off), [¶] Wgt-0 and Wgt-1 – the best weight set for ff03 and ff03/HB potentials respectively (no restriction on the sign of the weights), [‖] Wgt-2 and Wgt-R - the weight sets with allowed negative weights for dihedral (DIH), and for Wgt-2 also electrostatic (ELE, ELE1-4), and generalized Born solvation (GB) energies, [**] Wgt-3 - the weight set with all the weights positive, [§§] DIH – dihedral angle energy, [¶¶] VDW – van der Waals energy, [‖‖] VDW1-4 – short distance van der Waals energy (for atom pairs separated by less than four bonds), [***] ELE – electrostatic energy, [†††] ELE1-4 – short distance electrostatic energy (for atom pairs separated by less than four bonds), [‡‡‡] GB - generalized Born solvation energy, [§§§] SA – surface area dependent solvation energy, [¶¶¶] HB – hydrogen bond energy.

The electrostatic (ELE, ELE1-4) and GB energies are completely uncorrelated with TM-score and their weights are relatively small; therefore, their sign does not have much physical meaning. As expected, the weights of the energy terms that had initial low correlation coefficient (ELE, ELE1-4, GB) are relatively smaller than the weights of the terms showing larger initial correlation of energy with TM-score (VDW, SA, HB, VDW1-4, DIH).

The negative weight for the surface area energy (Table 3.1 and Table 3.3, SA) is partly an artifact of the optimization procedure and also reflects the weak average correlation of the SA energy with TM-score for our decoy set (CC = 0.36). The SA energy landscape is flat for a wide range of native similarity up to a RMSD from native > 8 Å, reflecting the low dependence of our decoy set on the radius of gyration and compactness of the decoys. Only in the near native region does the SA dependent energy component have a noticeable correlation with TM-score. The values of the SA energy are small compared to other energy components (roughly two orders of magnitude smaller than the electrostatic energy) and assigning it a negative weight probably helps balance some deficiencies of the correlation of the other energy terms. For a physical potential, we require a positive weight for the SA energy term, since it represents the hydrophobic energy, and it should energetically favor the transition from the unfolded conformation to a more globular one, not the opposite. Including more unfolded decoys in the force field optimization process should help to obtain a positive weight for the surface area dependent energy term.

Although the linear combinations of the components with some negative weights may produce a potential that correctly scores compact decoy structures, such a potential may not be useful for applications associated with the generation of the new structures (e.g. the refinement of the protein decoys). The most important future goal is to use the optimized force field for the refinement of protein decoys. For this purpose, we need a potential with the smallest number of negative weights but which still performs very well in decoy scoring and with a good energy - native-likeness correlation. Restricting the more weights to positive values decreases the performance of the potential, therefore we also chose the weight set that is a compromise between the number of negative weights and the performance, set ff03/HB Wgt-2. In Table 3.4 we show comparison of its performance with the best unrestricted ff03/HB Wgt-1 set and with the ff03/HB potential optimized keeping all weights positive, Wgt-3. Set Wgt-2 is the best performing weight set under requirement that the weights of the van der Waals (VDW), short-distance van der Waals (VDW1-4), surface area (SA), and hydrogen bond (HB) energies are positive (Table 3.3, Wgt-2). The negative weights for the electrostatic energy (ELE, ELE1-4) and GB solvation (GB) have no physical meaning, because these energy components are uncorrelated with native-likeness and their relative weights are small. We also allowed small negative values of the weights for the dihedral angles energy, because this energy is weakly anti-correlated with native-likeness and can be partially compensated by long and short-distance van der Waals interactions. As shown in Table 3.4, the performance of the Wgt-2 set is slightly worse compared to the Wgt-1 set; however, it is still significantly better that for the unoptimized force field (see Table 3.2, ff03/HB). The average correlation coefficient between the energy and TM-score, CC_{ave}, is 0.61, the percent of

proteins with a significant correlation coefficient, CC_{fr}, is 0.54, and the ability to indicate the native structure (TM_{fr}) and native cluster ($RMSD_{fr}$) remains very high, above 0.90. These results mean that using the ff03/HB Wgt-2 potential should allow for the refinement of decoy structures for about 54% of proteins.

Table 3.4 Comparison of scoring performance of the ff03/HB optimized force fields with different weight sets.

	Wgt-1*			**Wgt-2†**			**Wgt-3‡**		
	Train§	**Test¶**	**Set58‖**	**Train§**	**Test¶**	**Set58‖**	**Train§**	**Test¶**	**Set58‖**
CC_{ave}**	0.67	0.64	0.65	0.65	0.59	0.61	0.62	0.55	0.57
Z-$score_{ave}$††	2.59	2.19	2.29	2.49	1.86	2.02	2.12	1.49	1.65
CC_{fr}‡‡	0.60	0.65	0.64	0.73	0.47	0.54	0.60	0.42	0.47
TM_{fr}§§	1.00	0.86	0.90	1.00	0.86	0.90	0.73	0.72	0.72
$RMSD_{fr}$¶¶	1.00	0.88	0.91	1.00	0.88	0.91	0.80	0.79	0.79

* Wgt-1 - the best weight set (no restriction on the sign of the weights), † Wgt-2 - the weight set with allowed negative weights for dihedral (DIH), electrostatic (ELE, ELE1-4), and generalized Born solvation (GB) energies, ‡ Wgt-3 - the weight set with all the weights positive, § Train - training protein set, ¶ Test - testing protein set, ‖ Set58 – the entire set of 58 proteins, ** CC_{ave} - average correlation coefficient of the energy with TM-score, †† $Z\text{-}score_{ave}$ - average Z-score between native cluster and the remaining decoys, ‡‡ CC_{fr} - fraction of proteins with correlation coefficient of energy with TM score greater than 0.6, §§ TM_{fr} – fraction of proteins for which the lowest energy structure had the TM-score to the native state greater than 0.90, ¶¶ $RMSD_{fr}$ - fraction of proteins for which the lowest energy structure had the RMSD to the native state less than 2 Å.

We also analyzed the performance of the force field containing only positive weights. The best potential with all the weights positive, ff03/HB Wgt-3 performs slightly worse than the Wgt-1 and Wgt-2 potentials (see Table 3.4, column Wgt-3). The average correlation coefficient between the energy and TM-score, CC_{ave}, is 0.57, the percent of proteins with a significant correlation coefficient, CC_{fr}, is 0.47, and the ability to indicate the native structure (TM_{fr}) and native cluster ($RMSD_{fr}$) is still good, above 0.70. These results mean that using the ff03/HB Wgt-3 potential should allow for the refinement of decoy structures for about 47% of proteins. The weights for this potential are listed in Table 3.3, Wgt-3.

Comparison of the performance of the Wgt-1, Wgt-2, and Wgt-3 potentials is shown in Figure C.3 (Appendix C). The CC is still improved for the great majority of proteins, and the CC distribution is shifted toward the significant values for both Wgt-2 (Figure C.3, A' and B') and Wgt-3 (Figure C.3, A" and B"), compared with the unoptimized ff03/HB potential. The Z-score improved and is positive for all the proteins for both sets (Figure C.3, C' and C"). The scoring of the native structure and of the native cluster for the Wgt-2 is as good as for the Wgt-1 (Figure C.3, D' and E'), and becomes a bit worse for Wgt-3 (Figure C.3, D" and E").

3.3.7 Reduced optimized ff03/HB force field

As discussed in previous sections, the electrostatic and generalized Born energy terms have a very low correlation with TM-score and do not show specificity in recognizing the native structure. The magnitude of the electrostatic and generalized Born solvation energies is larger than the other energy components (roughly by an order of magnitude) and introduce a noisy uncorrelated background. During optimization, the

weights of these terms tend to decrease, resulting in the decrease of the background noise and increase of the relative contribution of the remaining energy terms to the total potential energy. For the purpose of protein structure refinement, which is our ultimate goal, it may be reasonable to turn off the electrostatics and generalized Born solvation energy. These components do not help drive the structure toward the native state and they are the most time consuming to calculate.

Following these arguments, we optimized the ff03/HB force fields with the weights set to zero for the electrostatic (ELE), short-distance electrostatic (ELE1-4), and generalized Born (GB) components of energy. As previously, the weights for the bond and angle energy were also set to zero. The optimization procedure was the same as described in the section "Force field optimization method". We chose the best performing weight set, Wgt-R (Table 3.3) with the requirement for positive weights for the VDW, VDW1-4, SA, and HB energy terms, allowing the dihedral energy (DIH) to have a small negative weight. In Table 3.5, we compare the performance of the reduced ff03/HB (Wgt-R) potential with the full ff03/HB (Wgt-2) force field, optimized under similar restrictions of positive weights for VDW, VDW1-4, SA and HB energy terms. For the Wgt-R, there is a slight decrease of performance compared to Wgt-2, visible in the change of average energy - TM-score correlation coefficient from 0.61 to 0.58 (CC_{ave}), the decrease of the percent of proteins with a significant CC from 0.54 for optimized ff03/HB to 0.43 (CC_{fr}), and slightly worse recognition of the native structure (TM_{fr}) and native cluster ($RMSD_{fr}$). Additional comparison of the performance of the Wgt-2 and Wgt-R is shown in Figure C.4 (Appendix C). Restricting weights to only positive values does not change the results significantly (results not shown).

Table 3.5 Comparison of scoring performance of the ff03/HB optimized (Wgt-2) and ff03/HB reduced optimized force fields (Wgt-R). Both potentials were optimized under similar conditions – allowing for a negative weight at the dihedral energy component (DIH). In Wgt-2, the electrostatic (ELE and ELE1-4) and generalized Born solvation (GB) energies also had negative weights, and in Wgt-R the corresponding weights are set to zero.

	ff03/HB optimized* Wgt-2			**ff03/HB reduced optimized† Wgt-R**		
	Train‡	**Test§**	**Set58¶**	**Train‡**	**Test§**	**Set58¶**
CC_{ave}‖	0.65	0.59	0.61	0.58	0.58	0.58
Z-$score_{ave}$**	2.49	1.86	2.02	1.69	1.51	1.56
CC_{fr}††	0.73	0.47	0.54	0.53	0.40	0.43
TM_{fr}‡‡	1.00	0.86	0.90	0.67	0.72	0.71
$RMSD_{fr}$§§	1.00	0.88	0.91	0.73	0.88	0.84

* Optimized ff03/HB potential, Wgt-2, † optimized ff03/HB reduced potential, Wgt-R (with electrostatic (ELE and ELE1-4) and generalized Born solvation (GB) energy components turned off), ‡ Train - training protein set, § Test - testing protein set, ¶ Set58 - entire set of 58 proteins, ‖ CC_{ave} - average correlation coefficient of the energy with TM-score, ** Z-$score_{ave}$ - average Z-score between native cluster and the remaining decoys, †† CC_{fr} - fraction of proteins with correlation coefficient of energy with TM score greater than 0.6, ‡‡ TM_{fr} - fraction of proteins for which the lowest energy structure had the TM-score to the native state greater than 0.90, §§ $RMSD_{fr}$ - fraction of proteins for which the lowest energy structure had the RMSD to the native state less than 2 Å.

The reduced force field should be able to refine structures for at least 43% of the proteins, find the native structure by the lowest energy criterion for over 70% of the proteins, and is less computationally demanding than the full potential. With electrostatics and generalized Born solvation energies turned off, the dominating weights are those for the short distance van der Waals (VDW1-4) and hydrogen bond (HB) energies (Table 3.3, Wgt-R).

3.4 Conclusions

In this work, we explored the applicability of a global optimization method based on a large set of protein decoy structures for many proteins to generate a funnel-shape of the energy to the native structure for an *AMBER* ff03 based, all-atom potential. Such potentials should enable the refinement of decoy structures toward the native state. We demonstrated that by including global energetic and structural data for a large set of protein decoy structures and by optimizing the relative weights of energy components of physics-based all-atom potential, it is possible to significantly improve the correlation of the energy with native-likeness and scoring of the native structure as the lowest in energy. Using such an approach to optimize the ff03/HB force field (the original Amber ff03 force field with an explicit hydrogen bond potential added), we improved the average correlation coefficient of the energy with TM-score from 0.25 (for the original ff03 potential) to 0.65, and the scoring of the native structure as the lowest in energy from 22% (for the original ff03 potential) to 90% of proteins, for a representative set of 58 proteins. Reaching an average correlation of 0.69 of energy with TM-score for the *TASSER* coarse-grained potential, developed earlier in our laboratory, allowed for the systematic refinement of the reduced protein models [12]. This gives as a reason to expect

that our optimized atomic potentials having a similar average energy - TM-score correlation will show systematic refinement ability.

We have also shown that the *DSSP* [94] hydrogen bond potential can significantly improve the correlation of the energy with native-likeness and the recognition of the native structure as the global energy minimum. Such a potential is sensitive to small changes of the orientation of the atoms that form a hydrogen bond. This results in a continuous increase of the energy of the structures as their hydrogen bonding deviates from a perfect pattern and a good correlation of the hydrogen bond energy with native-likeness, even in the region close to the native structure. Many well packed, but misfolded structures that have disturbed hydrogen bonding become higher in energy with respect to the native state.

For a large protein decoy sample, we observed that the electrostatic and generalized Born solvation energy components are uncorrelated with native similarity and do not show any specificity in recognizing the native state. The behavior of the electrostatic energy with native-likeness is protein-dependent, and there is no reason for the electrostatic energy to change monotonically with native similarity. In force fields, the "frozen" point charge approximation and absence of polarization additionally introduce unnaturally large fluctuations of the electrostatic energy, even for small changes of local geometry. The GB solvation energy also has an electrostatic character and suffers from the same large, nonphysical fluctuations as the electrostatic energy, caused by the point charge approximation. The solvation energy is usually favorable for extended structures and for some proteins is weakly anti-correlated with native similarity, as the structure becomes more compact and less solvated. The electrostatic and

generalized Born solvation energies comprise a noisy uncorrelated background to the other energy components. As a result of optimization, the weights of these energy components decrease, suggesting the limited role of electrostatic energy and electrostatic component of solvation in directing the structure towards the native state. In contrast, a stronger initial correlation of energy with native-likeness is observed for the van der Waals and the hydrogen bond energy. The weights of these energy components become relatively larger after force field optimization.

The dihedral energy (DIH) interaction appears to be weakly anti-correlated with native-likeness, which results in a negative, but small weight of this energy component in some of our optimized potentials. This anti-correlation is in agreement with our earlier observations of the tendency of the ff03 force field to distort the dihedral angles from their gas phase equilibrium values for short helices and strands of polypeptides (as compared to the quantum mechanical calculations). These two results suggest that the dihedral energy term may require reoptimization.

Since the electrostatic and generalize Born solvation energy components acquire small weights during optimization, we explored the use of a reduced potential with the electrostatic and GB solvation terms turned off. The scoring performance of the optimized reduced ff03/HB force field (Wgt-R) is worse than the performance of the optimized full ff03/HB (Wgt-1) by 5% for the average correlation coefficient of energy with TM-score, 20% for the percent of proteins with a significant correlation coefficient, and 21% for the percent of proteins for which the native structure has the lowest energy. Therefore, the loss of performance of the optimized reduced potential compared to the full optimized ff03/HB force field is not very large, and for 43% of proteins, the

correlation coefficient of energy with TM-score is larger than 0.60, allowing correct decoy scoring. The reduced optimized potential is significantly better than the full unoptimized ff03 and ff03/HB force fields.

The ultimate goal of global optimization of the force fields is not only the correct scoring of protein decoys but also the refinement of low-resolution models. In Chapter 4, we describe our results in the refinement tests on the newly developed ff03/HB reduced optimized potential.

CHAPTER 4

REFINEMENT OF PROTEIN STRUCTURES USING AN OPTIMIZED, PHYSICS-BASED ALL-ATOM FORCE FIELD

4.1 Introduction

Significant progress has been made in the field of protein structure prediction [1-3]. Contemporary methods are able to assemble the correct topology for a large fraction of protein domains. But even such approximately correct models typically vary in the structural similarity to the native state and range from 1 to about 6 Å RMSD (root mean square deviation) from native. Models with resolution of 1-2 Å have a reliability comparable to experimentally obtained structures and can be used in a broad range of applications, including studies of reaction mechanisms, functional annotation, drug design and virtual ligand screening. For low-resolution models (3 - 6 Å away from the native), the spectrum of useful applications is much narrower [2]. Structure prediction methods use a coarse-grained representation of proteins to simplify the search problem, and it is possible that the structural details are necessary to improve the packing of the protein core and the quality of the model. A tempting approach is to use an all-atom detailed protein representation in the endgame of structure prediction, but despite efforts, all-atom refinement has so far seen little success. There have been reports of single examples of successful refinements [48-50], with the best improvement of 2 Å [51]. Also a few studies reported refinement benchmark [15,53,55] for a set of proteins. Such, more comprehensive results show that even though single examples of refinement do occur for

some protein models, the methods are far from routine, and usually most models deteriorate instead of improve. Also, due to the high computational cost, the largest benchmark set contained only 15 proteins.

In protein structure prediction and refinement, the challenge is two fold: the first problem is the conformational search and the second is the inaccuracies of the energy function. The failure to refine protein models may be attributed to the problem with generating native-like structures during the search. But sampling is guided by energy, and the potential function has to drive the search towards native-like regions. The energy should be able to find the native structure among decoys and it should have a correlation with native similarity. In our previous work [96], described in Chapter 3, we explored the possibility of creating a funnel-like shape for the *AMBER* [74] potential by global optimization of the weights of particular energy components. The optimized force field had a significant correlation with native similarity and was able to recognize the native conformation among decoy structures for a large fraction of proteins. Here, we test the refinement ability of the newly derived potential. Using 47 representative proteins and a diverse set of compact all-atom decoys, we obtain improvement of the quality of the models (as measured by TM-score [72]) in 70% of the cases when the lowest energy structure from the refinement run is compared with the starting model. Only 18% of all decoys deteriorate relative to the native structure, and 12% do not change. Moreover, only for three proteins, did we observe improvements of less then 50% of the decoys. Our study presents the first systematic refinement of protein models and the most comprehensive benchmark for all-atom model refinement.

4.2 Methods

4.2.1 Conformational search method

To search the conformational space of proteins in the refinement procedure, we used our newly developed *A-TASSER* program [96]. *A-TASSER* (for *atomic-TASSER*) represents the protein at atomic detail and employs the Replica Exchange Monte Carlo (REMC) [88,89] search method with a Parallel Hyperbolic Sampling (PHS) acceptance criterion [90] to reduce higher energy barriers. *A-TASSER* uses three types of moves that change only the torsional angles of the molecule: local "fixed end" moves [91], end moves, and the side chain moves. The details of *A-TASSER* are described in Chapter 3.

4.2.2 Force field

The potential energy function employed in this study to refine protein models is calculated according to Equation 1:

$$E_{FF03/HB/R} = w_{DIH}E_{DIH} + w_{VDW}E_{VDW} + w_{VDW1-4}E_{VDW1-4} + w_{SA}E_{SA} + w_{HB}E_{HB} \qquad \text{Eq.1}$$

In Equation 1, the following abbreviations and symbols are used: E – denotes the energy, w – is the weight of a given energy component, DIH – dihedral term, VDW – van der Waals component, VDW1-4 – van der Waals energy for atom pairs separated by less than four bonds, SA – surface area dependent term (hydrophobic component of solvation), and HB hydrogen bond term. The E_{DIH}, E_{VDW}, E_{VDW1-4}, and E_{SA} energy terms are identical with those in ff03 Amber force field. The E_{HB} hydrogen bond energy was implemented following the *DSSP* approach [94] and is described in the previous chapter (Chapter 3). The weights of the energy terms, w (Table 4.1), were adjusted using a global optimization

method [92] for a large set of decoy structures of the representative 58 protein set [96]. The optimization procedure is aimed at maximizing the correlation of the energy with TM-score [72] and maximizing the energy gap between the native state and the decoys. The force field used in this study has an average correlation coefficient of energy with TM-score of 0.59 and ranks structures with TM-score larger than 0.9 (native-like) as the lowest in energy for 72% of proteins [96].

Table 4.1 Relative* weights of energy components in the optimized force fields.

	BOND	ANG	DIH	VDW	VDW 1- 4	ELE	ELE 1- 4	GB	SA	HB
ff03/HB/R optimized	0	0	-0.42	1.00*	4.33	0	0	0	0.51	4.26

*All weights were scaled so that the weight for van der Waals energy is equal to 1. Abbreviations used for particular energy components: BOND – bond, ANG – bond angle, DIH – dihedral angle, VDW – Van der Waals, VDW1-4 – Van der Waals term for atom pairs separated by less then four bonds, ELE – electrostatics, ELE1-4 – electrostatics for atom pairs separated by up to four bonds, GB – generalized Born approximation to polar solvation, SA – non-polar, surface area dependent solvation, HB – hydrogen bond.

The optimized force field does not include electrostatic and generalized Born solvation [29] energy terms. Such a reduction did not substantially compromise the scoring performance of the force field, and resulted in a much shorter time for the energy calculation. The decision to exclude the electrostatic and GB components was based on our previous findings [96] that these energy terms in the ff03 force field were large and uncorrelated with native similarity of protein models. Therefore, they do not drive the conformational search towards the native state. The bond and angle energy components

are also excluded from the potential, because the sampling method keeps the bonds and valence angles unchanged.

4.2.3 Protein set and starting decoy structures

We tested our method on 47 proteins, a subset of a previously prepared [72] comprehensive benchmark set, which includes 1489 test proteins and covers the PDB library [59] with lengths from 41 to 200 residues at 35% sequence identity. The 47 proteins are also a subset of the 100-set of proteins described in Chapter 2. The chosen proteins span the lengths from 54 to 123 residues and represent different secondary structural groups. The list of proteins can be found in Table B.2 (Appendix B). Among these 47 proteins, eight (marked in Table B.2) were a part of the training set used in the optimization of the force field [96] and they were excluded for most analyses to avoid any possible memorization effects. Figure 4.1 shows results for 47 proteins; all other results include only the 39 testing proteins. For each protein, we randomly chose 100 decoys from the force field optimization decoy set such that they span the range of Cα RMSD to the native structure from 0 to 8 Å. These 100 decoys per protein and the native structures in all-atom representation were starting models in our refinement benchmark.

4.2.4 Refinement protocol

For each decoy, we ran an *A-TASSER* search consisting of 1000 swaps between replicas, and 200 steps of PHS at each replica between swaps. From each decoy trajectory, the lowest or the best of the 5 lowest energy structures were selected for analysis as the refinement results. No clustering was used in decoy selection.

4.3 Results and Discussion

4.3.1 Refinement of protein decoys

During the refinement, the TM-score and RMSD improve for a majority of the decoys. In Figure 4.1 the TM-score (A) and Cα RMSD (B) to the native structure of the lowest energy decoy from each refinement trajectory is compared with the initial decoy TM-score (RMSD). The result includes all 47 proteins used in this study and all their decoys. For TM-score, we observe a more pronounced improvement compared to decoy deterioration compared to that for the RMSD, which reflects the force field optimization procedure that maximized the correlation coefficient of the energy with TM-score, not with RMSD. Sometimes the improvements of TM-score may cause an increase of RMSD from the native structure, e.g. when the core of a protein is improved at the cost of moving a protein tail farther from the native state. Below a 1 Å RMSD (or above 0.9 TM-score), this force field cannot differentiate among structures. This effect is visible in Figure 4.1, the native structure (TM-score close to 1) drifts away from the initial structure on average by ~0.1 TM-score or 1 Å in RMSD. This drift determines the resolution of the force field, which is about a 1 Å Cα RMSD to the native state.

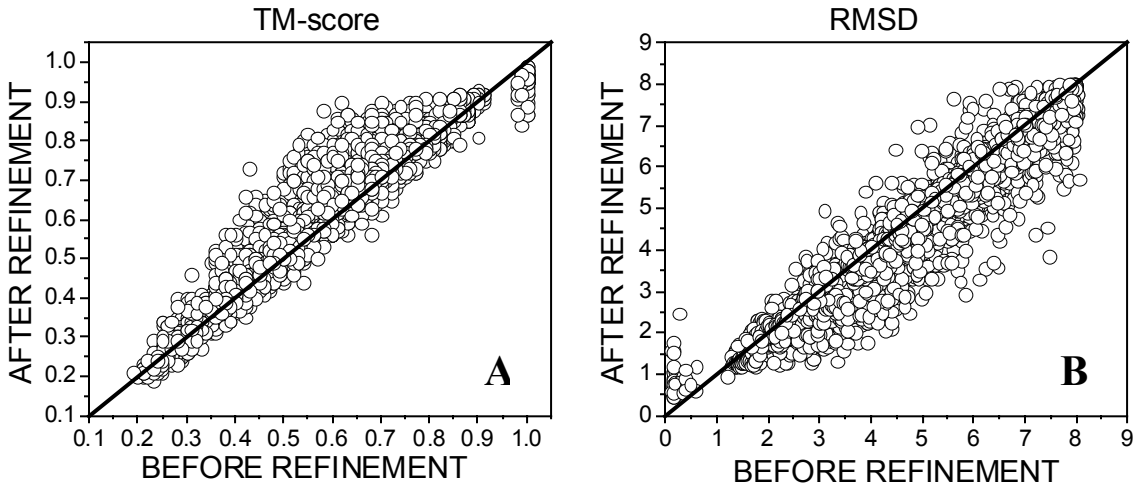


Figure 4.1 TM-score (A) and RMSD (B) from the native structure are plotted for each model and the native before and after refinement. The structure after refinement is the lowest energy conformation from the refinement trajectory. All models for 47 proteins are presented.

On average, over the whole range of native similarity, 70% of decoys improve their TM-score with respect to the initial structure, 18% get worse and 12% do not change. When RMSD is used as native similarity measure, the changes are: a 70% improvement, a 28% deterioration and 2% do not change respectively. We then consider separately the changes of starting decoys from different native similarity bins. Over the entire range of TM-score and RMSD of starting decoys, the fraction of decoys that improve with respect to the initial model during refinement, dominates over those that deteriorate. Deterioration only dominates in the close near-native region (1-0.9 TM-score and 0-1 Å RMSD), due to the resolution of the force field.

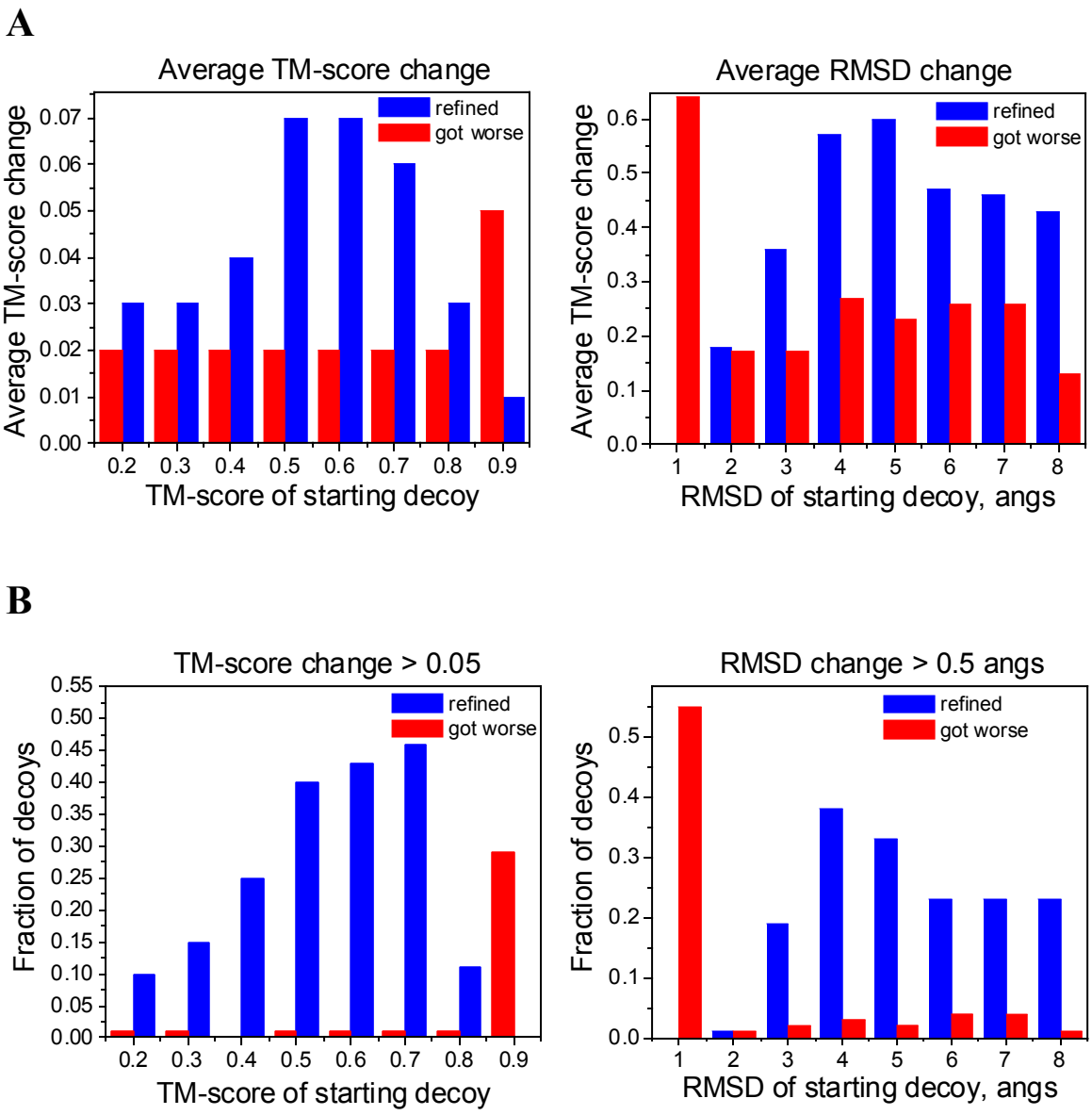


Figure 4.2 Structural changes of the decoys during refinement with respect to the native structure, in different native similarity bins. A: Average TM-score and RMSD changes per bin, B: Fraction of decoys that changed by more than 0.05 TM-score or 0.5 RMSD. Blue denotes improvement, and red deterioration of the structure.

Figure 4.2 shows the average (A) and the significant (B) changes in decoy quality during refinement in different native similarity bins. On average, improvements are larger than

deteriorations over the whole native similarity range, except for the native bin (Figure 4.2 A). Also the native state is quite stable during the simulations, and the structures deteriorate by only about 0.05 TM-score units, or 0.65 Å in RMSD. Figure 4.2 B shows the fraction of decoys that change significantly during refinement by more than 0.05 in TM-score or more than 0.5 Å in RMSD. The fraction of decoys that significantly deteriorate is negligible, except for the region close to the native structure; in contrast significant improvements are observed for as much as ~40% of decoys in the TM-score range of 0.5-0.7.

In Figure 4.3, we show the distribution of the fraction of decoys that improve and deteriorate in the set of 47 proteins. For most proteins, more than 50% of the decoys improve. Only four proteins, 1a19, 1b9wA, 1c1yB, and 1dt4A, have less than a 50% improvement. Among these, only two (1b9wA and 1dt4A) had more decoy deterioration than improvement. Protein 1b9wA has 5 disulfide bonds, and there is no specific disulfide bond potential in our force field; this may be one of the reasons for the inferior results for this particular protein. There are nine proteins in the set that contain disulfide bonds (1a43, 1aazA, 1bunB, 1bvnT, 1cc7A, 1dtdB, 1f94A, 1b9wA, 1cbp). The refinement results for their decoys are on average worse than the average results for the remaining proteins (61% improvement, 26% deterioration, for the proteins with disulfide bonds, compared to 70% improvement and 18% deterioration for the remaining proteins). This indicates the need to include an additional disulfide bond potential. It is especially important for small proteins whose fold is mainly held together by S-S bridges. Based on the above results, we conclude that our optimized force field enabled significant and

systematic refinement of protein structures with respect to the initial decoys and is driven only by energetic criteria.

Examples of refined structures are shown in Figure 4.4. The largest observed improvement in TM-score was 0.30 (from 0.53 to 0.85 for 1b07A).

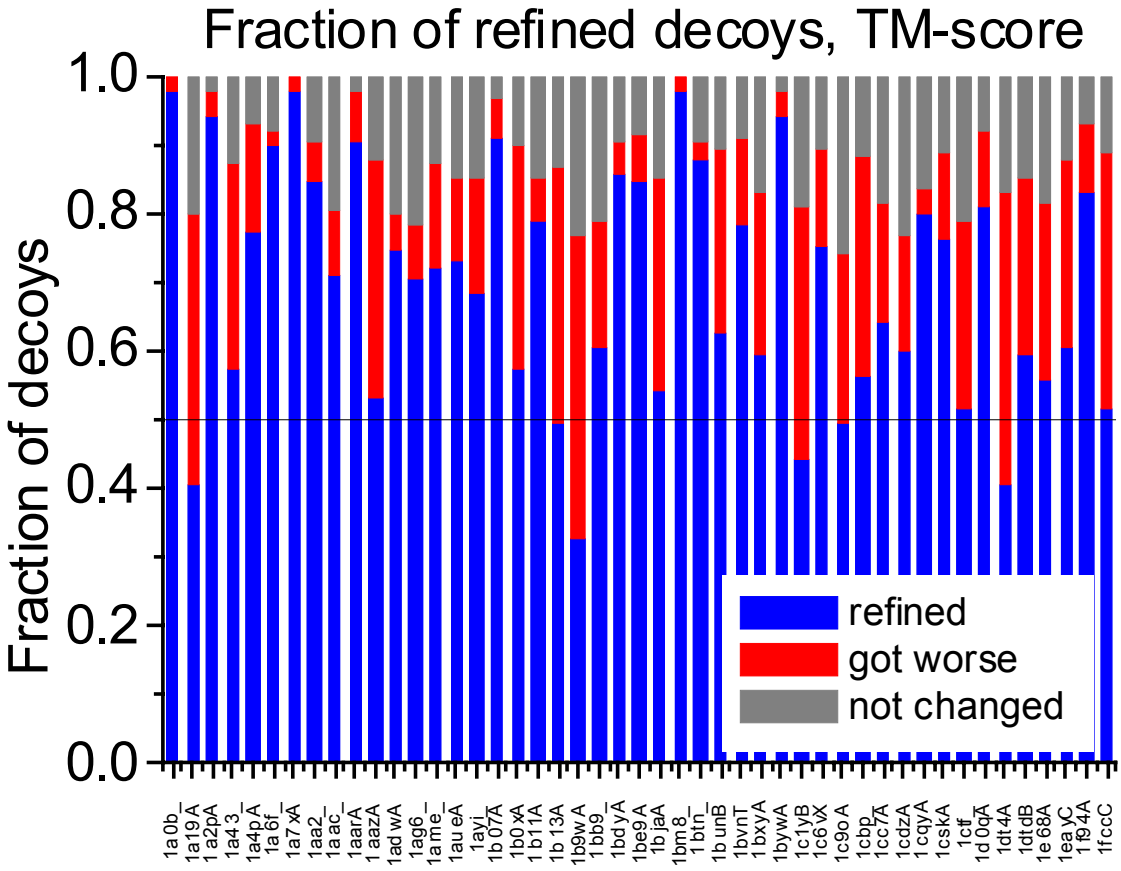


Figure 4.3 Fractional changes of decoys for each of the 47 proteins. Blue denotes refinement, red – deterioration, gray – no change in the TM-score with respect to the native.

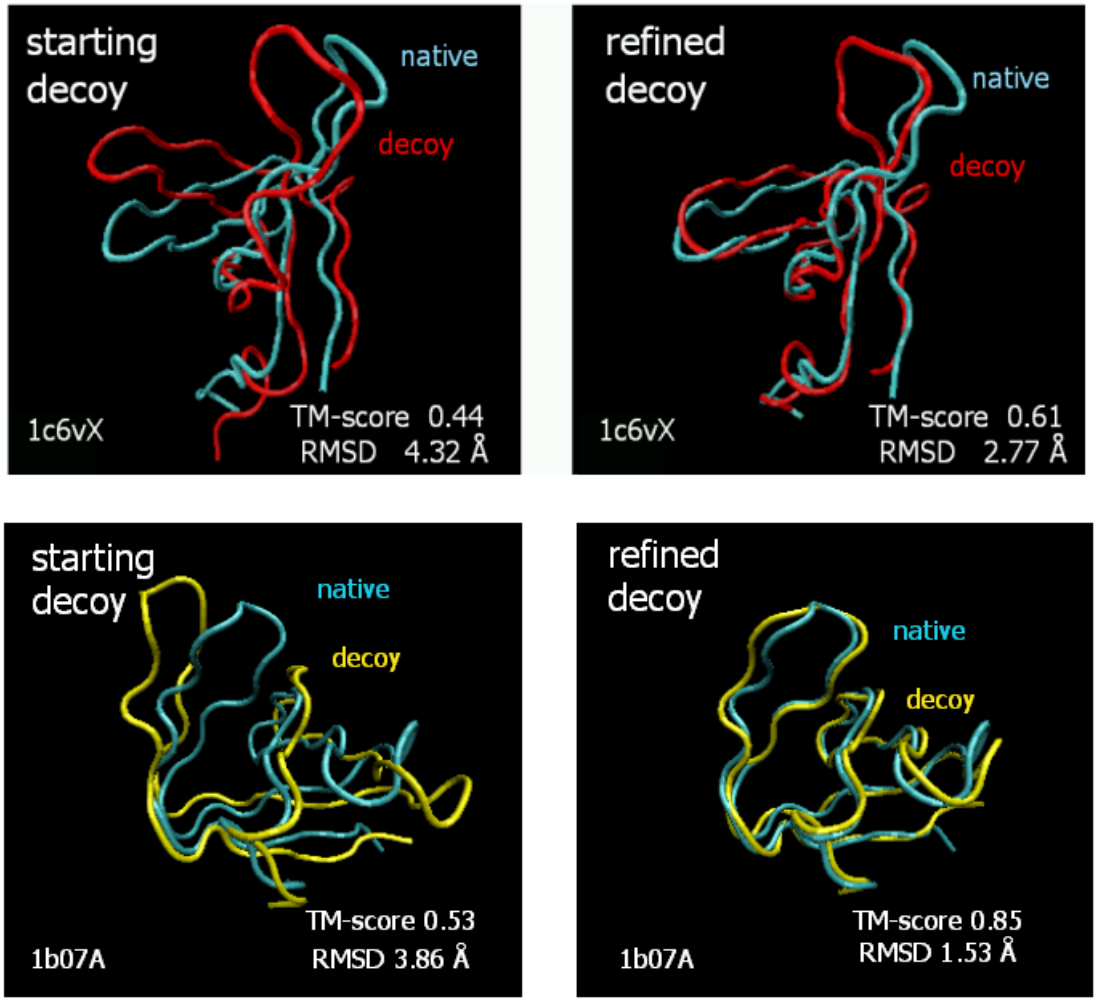


Figure 4.4 Examples of decoy refinement for proteins 1c6vX and 1b07A. The refined decoy is the lowest energy structure from the refinement trajectory.

4.3.2 TM-score and RMSD of the lowest energy structure to the native state

Previous analysis showed the refinement performance within each decoy trajectory; the lowest energy structure was chosen from each decoy trajectory and separately compared with the starting decoy. In this section, we analyze the TM-score and RMSD to the native state of: a) the lowest energy decoy, and b) the best out of five

lowest energy decoys among the entire ensemble of refined decoy structures for each protein. In Figure 4.5, we show the fraction of proteins for which the lowest energy refined structure had a TM-score larger than (RMSD lower than) the specified threshold value. In this analysis, we consider only the 39 proteins that were not previously used in the optimization of the force field [96]. When the trajectory of the native structure is included, for 79% of proteins, the lowest energy structure has a TM-score to the native state above 0.7 and 82% of proteins have a Cα RMSD to native below 3.5 Å (Figure 4.5, black bars). Structures coming from the native trajectory are usually better packed than decoys and therefore favored by energy. A more stringent test of the force field and a better reflection of the real prediction conditions is to check the native-likeness of the lowest energy structures when the native decoys are excluded from the ensemble. Under such conditions, 59% of proteins[1] have lowest energy structures with a TM-score to the native state above 0.70, and 66% of proteins[1] have their lowest energy structure with a Cα RMSD to the native below 3.5 Å (Figure 4.5, gray bars). We additionally consider the best structure (highest TM-score or lowest RMSD to the native state) out of the five lowest energy conformations. With the native decoys included in such analysis 87% proteins have the best-of-five structure above a TM-score of 0.70, and 90% of proteins have the best-of-five structure below 3.5 Å RMSD. When the native decoys are excluded, 68% of proteins have the best-of-five structure above a TM-score of 0.70, and 77% of proteins have the best-of-five structures below 3.5 Å respectively. Based on the above

---

[1] The statistics was prepared for the subset of proteins that had decoys within given range of RMSD and TM-score.

results, we conclude that the ability of the force field to find the native structure among decoys is quite high. Also high is the chance to pick a good structure (that has an RMSD below 3.5 Å from the native state) among decoys, when the native structure is not present in the ensemble.

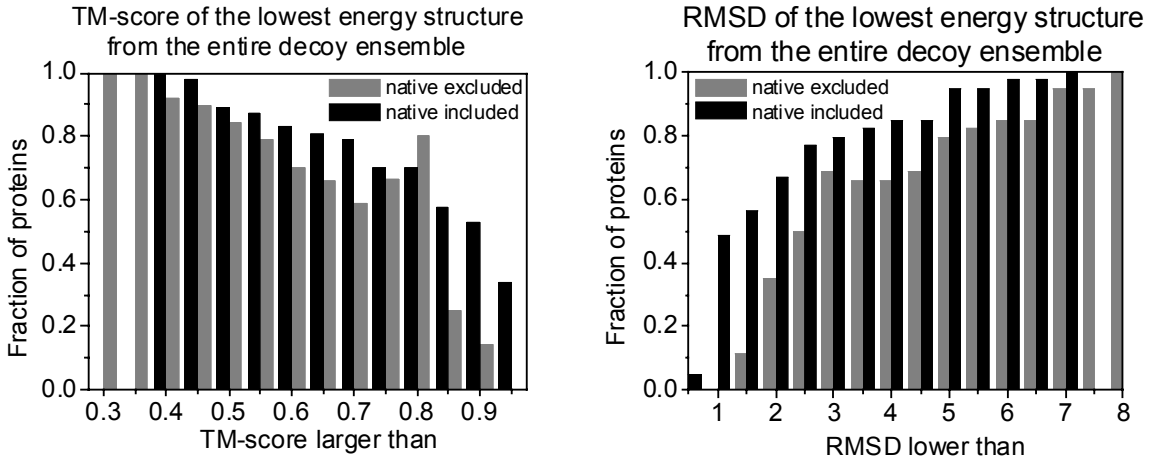


Figure 4.5 The fraction of proteins for which the lowest energy refined structure is within given native similarity threshold value (measured by TM-score and RMSD). Black bars: native trajectory included in the calculation, gray bars: native trajectory excluded. Only 39 proteins not used previously in the optimization of the force field [96] are included.

4.3.3 Correlation of the energy with native similarity measured by TM-score

The ability of a force field to refine a model and select the native or close to the native structures using the energy as the selection criterion is related to the correlation of the energy with native similarity. The force field used in this study had an average correlation coefficient of energy with TM-score of 0.58 after global optimization (as calculated for a large decoy set of 43 testing proteins) [96]. Moreover, for 40% of the tested proteins the correlation coefficient was significant, above 0.60. The decoys used in the optimization procedure and in the calculation of the correlation coefficients were generated in a different force field (ff03 *AMBER* potential). During the refinement, a broad search with the optimized potential is conducted that may reveal a different energy landscape. In particular, the high correlation coefficient of energy with native-likeness may only be an artifact of optimization and may be lost during a thorough conformational search with the new potential. To explore this issue, we calculated the correlation coefficient (CC) between the energy and TM-score for the structures generated during conformational refinement. Again, only the 39 testing proteins were used. The resulting average value of CC was 0.59, with 46% of proteins having a CC above 0.60, which corresponds well to previously obtained values after force field optimization [96], but without use of the force field to drive the conformational search. Therefore, we can conclude that during the conformational search employed for the purpose of this study, the good characteristics of our optimized energy landscape are preserved.

CHAPTER 5

CONCLUSIONS

5.1 Summary of the results

The presented thesis research focused on the problem of the refinement of low-resolution protein models to higher resolution as a part of a protein structure prediction procedure. In Chapters 2 and 3, we explored the possibility of using contemporary physics-based all-atom force fields for model refinement. The results revealed that the native structure is not the lowest conformation for the majority of proteins in the tested potentials, and therefore the force fields cannot drive the conformational search toward the native. Indeed, most models drift farther away from the native structure during the search.

Guided by the test results, we then attempted to globally optimize the best performing, *AMBER* ff03 potential to create a funnel-like shape for the energy surface. We changed the relative weights of particular components of the ff03 force field such that the final energy function has a global minimum in the native state and an improved correlation with native-likeness. Additionally, we supplemented the original force field with an explicit hydrogen bond potential. This resulted in an optimized force field with a significant correlation coefficient between energy and native similarity. Also, the native structure had the lowest energy among alternative conformations for most tested proteins.

Finally, we tested the newly developed force field in refinement. For a diverse set of decoys of 47 proteins, we observed refinement for 70% of the structures. Only 18% of the decoys deteriorated during the test, and 12% did not change with respect to the native

structure. Moreover, the most significant structural changes mostly resulted in more native-like structures and the average improvement was larger than the average deterioration. Such a systematic refinement has never before been reported. These results are extremely promising in the context of high-resolution structure prediction.

5.2 Future Work

The developed refinement tools are now being incorporated into an automatic protein structure prediction pipeline. In the next step, we will formulate a confidence score for successful refinement. Such a score will most likely be based on refinement convergence criteria (clustering).

The current formulation of our optimized force field does not include electrostatic interactions. This may in principle lead to the generation of wrong conformations, with charged residues buried in the protein interior. Even though our procedure is designed for searches that start from conformations with an already assembled overall topology, we plan to further explore the issue. Should it be necessary, we will develop a simplified interaction potential (e.g. short-range) for ionizable residues.

As tested on the set of refined decoy structures, our potential did not lose the desired correlation with native similarity after long search. We will additionally test if these characteristics are preserved when the search time is further extended. After a much longer search, deeper energy minima may be found that are away from the native structure, as we have observed for the unoptimized *AMBER* force field. In such a case, we can iterate the optimization procedure, and reoptimize the potential.

APPENDIX A

DEFINITIONS

Definition of the TM-score [72] (Template Modeling score):

$$TM-score = Max\left[\frac{1}{L_N}\sum_{i=1}^{L_T}\frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}\right]$$

$$d_0 = 1.24\sqrt[3]{L_N - 15} - 1.8$$

L_N is the length of the native structure, L_T is the length of the aligned residues to the template structure, d_i is the distance between the i-th pair of aligned residues and d_0 is a scale to normalize the match difference. 'Max' denotes the maximum value after optimal spatial superposition. The value of the TM-score always lies between (0,1], with better templates having higher TM-score. The d_0 distance is an empirical variable introduced to eliminate the protein size dependence in the TM-score.

APPENDIX B

SUPPLEMENTARY TABLES

Table B.1 (CHAPTER 3) Correlation coefficients of the energy with TM-score for individual proteins from Set58 before optimization (ff03, ff03/HB) and after optimization (ff03 optimized, ff03/HB optimized) force fields; [§] correlation coefficients for original unoptimized ff03 potential, [¶] correlation coefficients for unoptimized ff03/HB potential (ff03 with added hydrogen bond potential), [‖] correlation coefficients for optimized original ff03 potential (weight set Wgt-0), [**] correlation coefficients for optimized ff03/HB potential (ff03 with added hydrogen bond potential, weight set Wgt-1), [†] average (over 58 proteins) correlation coefficients (standard deviation in parentheses), proteins 1-15 constituted the first training protein set (Train1), proteins 16-30 constituted the second training set (Train2).

	PDB ID	Number of residues	Secondary structure	ff03[§] CC	ff03/HB[¶] CC	ff03 optimized[‖] Wgt-0 CC	ff03/HB optimized[**] Wgt-1 CC
1	121p_	166	α/β	0.01	0.07	0.57	0.67
2	1a19A	89	α/β	0.02	0.13	0.48	0.86
3	1abmA	198	α/β	0.13	0.20	0.41	0.56
4	1afcA	127	β	-0.01	0.09	0.49	0.67
5	1ahq_	133	α/β	0.17	0.21	0.84	0.46
6	1b0xA	72	α	-0.01	0.07	0.42	0.67
7	1bb9_	83	β	0.14	0.26	0.30	0.72
8	1bjaA	95	α	0.15	0.25	0.43	0.81

Table B.1 continued

9	1bxyA	60	α/β	0.12	0.28	0.94	0.89
10	1c0fS	127	α/β	0.12	0.15	0.67	0.25
11	1c4zD	144	α/β	0.48	0.51	0.77	0.67
12	1c9oA	66	β	0.44	0.50	0.87	0.74
13	1cskA	58	β	0.34	0.44	0.60	0.80
14	1d0qA	102	α/β	0.16	0.20	0.81	0.81
15	1elkA	153	α	0.07	0.17	0.86	0.78
16	1a0k_	130	α/β	0.34	0.38	0.73	0.59
17	1aa2_	108	α	0.24	0.28	0.91	0.54
18	1adwA	123	β	-0.01	0	0.40	0.26
19	1ag6_	99	β	0.31	0.35	0.46	0.57
20	1aueA	92	α	-0.05	0.14	0.52	0.90
21	1b07A	58	β	0.34	0.40	0.65	0.65
22	1b1bA	140	α	0.02	0.07	0.53	0.44
23	1be9A	99	β	0.19	0.22	0.88	0.48
24	1bm8_	99	α/β	0.90	0.91	0.90	0.89
25	1bz4A	144	α	0.50	0.58	0.03	0.86
26	1c1yB	77	α/β	0.27	0.35	0.72	0.77
27	1c6vX	55	β	-0.44	-0.36	0.65	0.70
28	1cdzA	96	α/β	-0.13	-0.01	0.76	0.75

Table B.1 continued

29	1ctf_	68	α/β	0.27	0.38	0.92	0.76
30	1fccC	56	α/β	0.14	0.29	0.57	0.84
31	1ayi_	86	α	0.18	0.30	0.87	0.75
32	1eayC	67	α/β	-0.08	0.04	0.69	0.69
33	1a2pA	108	α/β	0.38	0.41	0.49	0.50
34	1a33_	174	β	0.34	0.39	0.59	0.67
35	1a3aA	145	α/β	0.05	0.14	0.48	0.61
36	1a3k_	137	β	0.30	0.38	0.51	0.60
37	1a3s_	158	α/β	0.22	0.28	0.48	0.61
38	1a3z_	150	β	0.27	0.33	0.49	0.58
39	1a44_	185	β	0.23	0.25	0.56	0.48
40	1a45_	173	β	0.12	0.14	0.47	0.31
41	1a4pA	92	α	0.12	0.17	0.14	0.28
42	1a6f_	113	α/β	0.15	0.25	0.93	0.90
43	1a6jA	150	α/β	0.08	0.15	0.51	0.53
44	1a7xA	107	α/β	0.87	0.88	0.96	0.91
45	1aac_	105	β	0.25	0.29	0.11	0.32
46	1aep_	153	α	0.27	0.35	0.87	0.80
47	1bdyA	123	β	0.03	0.11	0.08	0.42
48	1befA	177	β	0.01	0.07	0.71	0.55

Table B.1 continued

49	1btn_	106	β	0.15	0.21	0.42	0.47
50	1bywA	110	α/β	0.72	0.74	0.87	0.83
51	1c02A	166	α	0.64	0.67	0.86	0.74
52	1c25_	161	α/β	0.27	0.30	0.42	0.50
53	1a0b_	117	α	0.83	0.86	0.91	0.87
54	1em9A	147	α	0.87	0.88	0.95	0.88
55	1aarA	76	α/β	0.55	0.58	0.67	0.70
56	1ame_	66	β	0.84	0.84	0.94	0.86
57	1b11A	113	β	0.24	0.31	0.56	0.66
58	1b13A	54	β	0.38	0.39	0.05	-0.02
	Average[†]			**0.25 (0.25)**	**0.31 (0.25)**	**0.62 (0.25)**	**0.65 (0.19)**

Table B.2 (CHAPTER 4) List of the 47 proteins used in refinement tests (proteins marked with ^{T} were used in the training set in the optimization of the force field [96]).

#	PDB ID	Number of residues	SCOP class
1	1a0b_	117	α
2	1a19A ^{T}	89	α/β
3	1a2pA	108	α+β
4	1a43_	72	α
5	1a4pA	92	α
6	1a6f_	72	α+β
7	1a7xA	107	α+β
8	1aa2_	108	α
9	1aac_	105	β
10	1aarA	76	α+β
11	1aazA	87	α/β
12	1adwA	123	β
13	1ag6_	99	β
14	1ame_	66	β

Table B.2 continued

15	1aueA	92	α
16	1ayi_	86	α
17	1b07A	58	β
18	1b0xA^{T}	72	α
19	1b11A	113	β
20	1b13A	54	small
21	1b9wA	89	small
22	1bb9_^{T}	83	β
23	1bdyA	123	β
24	1be9A	115	β
25	1bjaA^{T}	95	α
26	1bm8_	99	α+β
27	1btn_	106	β
28	1bunB	61	small
29	1bvnT	71	β
30	1bxyA^{T}	60	α+β
31	1bywA	110	α+β
32	1c1yB	77	α+β
33	1c6vX	55	β
34	1c9oA^{T}	66	β

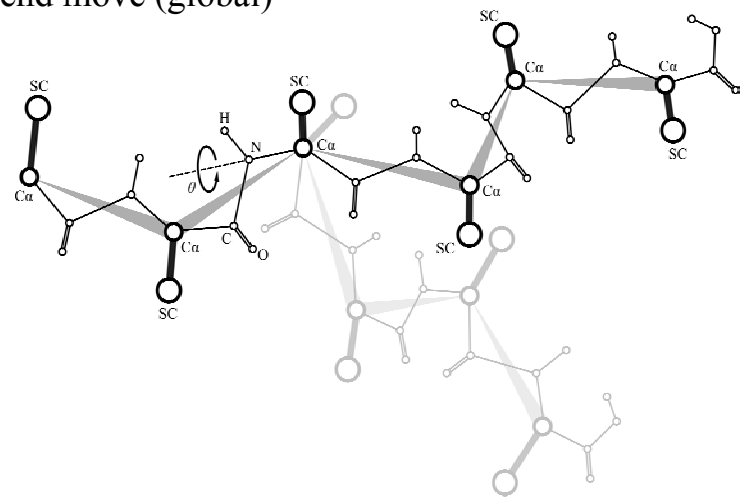
Table B.2 continued

35	1cbp_	86	β
36	1cc7A	72	α+β
37	1cdzA	96	α/β
38	1cqyA	99	β
39	1cskA^{T}	58	β
40	1ctf_	68	α+β
41	1d0qA^{T}	102	small
42	1dt4A	73	α+β
43	1dtdB	61	small
44	1e68A	70	α
45	1eayC	67	α+β
46	1f94A	63	small
47	1fccC	56	α+β

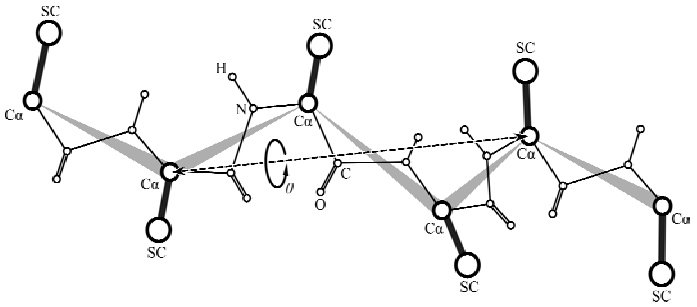
APPENDIX C

SUPPLEMENTARY FIGURES

A end move (global)



B fixed-end move (local)



C side chain move

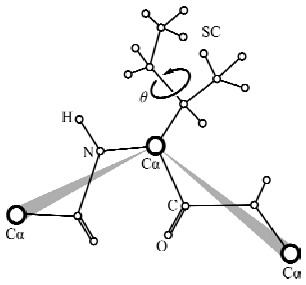


Figure C.1 (CHAPTER 3) Schematic representation of the moves used by the *A-TASSER* conformational search program. The dashed line and the arrow represent the rotation axis, with θ the rotation angle. A - end move, the rotation of a φ or ψ backbone angle that involves 1-5 residues at the ends of the molecule, B - local "fixed-end" move that is conducted along the axis connecting two Cα atoms and involves 2-12 residue fragments in the molecule's interior, C - side chain move involves any torsion angle of a given side chain.

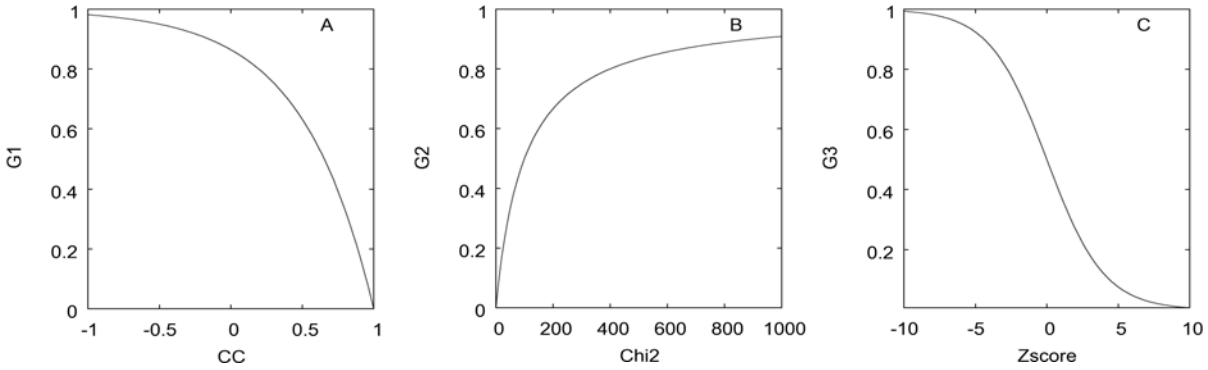


Figure C.2 (CHAPTER 3) Graphical representation of the components of the target optimization function F, A - G1 (Eq.3), function of the correlation coefficient of the energy with TM-score, B – G2 (Eq.4), function of the χ^2 value of the linear fit of energy versus TM-score dependence, C – G3 (Eq.5), function of Z-score between energy of the native and non-native decoys clusters.

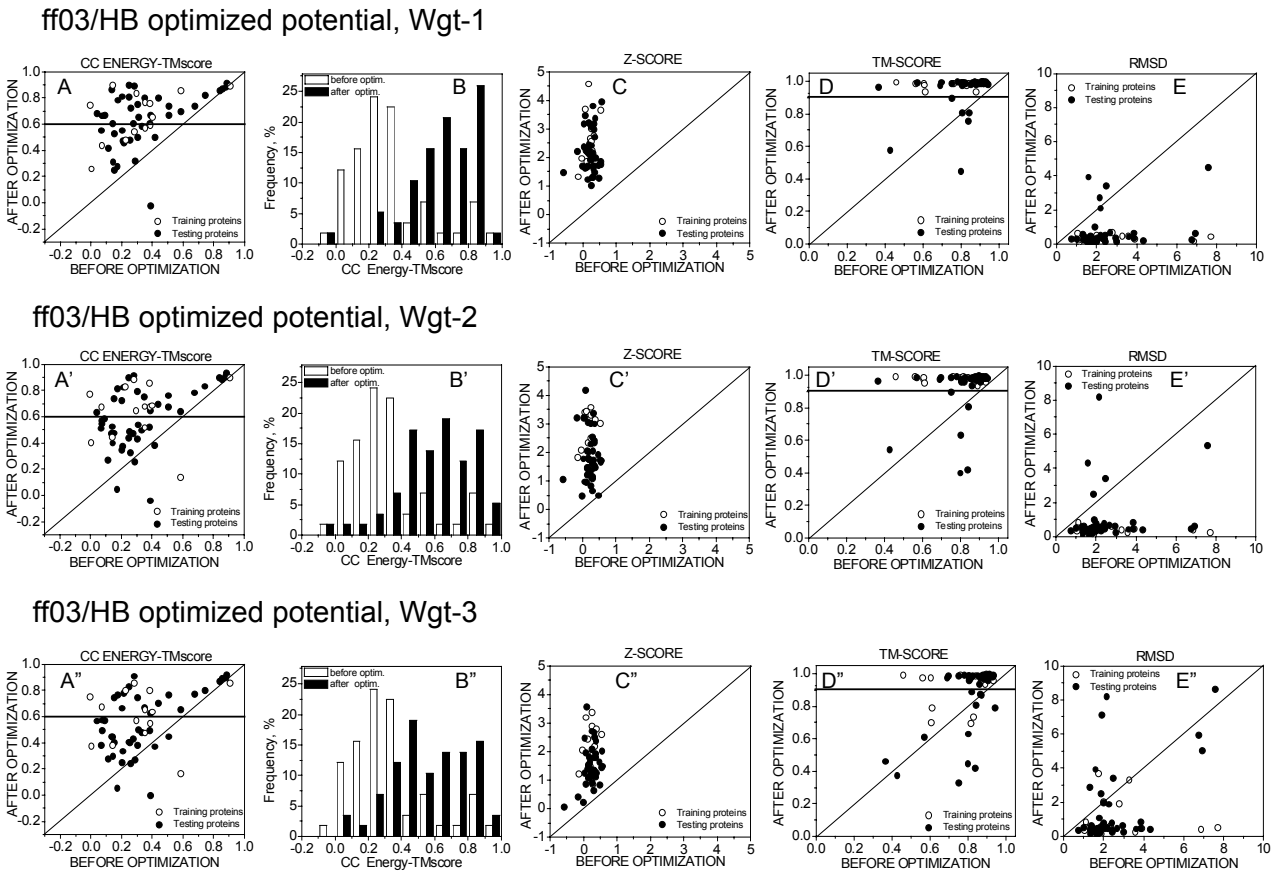


Figure C.3 (CHAPTER 3) Comparison of the scoring performance of the optimized ff03/HB force fields with different weights for the set of 58 proteins (Set58), A-E – the best weight set (Wgt-1), A'-E' – the weight set with allowed negative weights for dihedral (DIH), electrostatic (ELE, ELE1-4), and generalized Born solvation (GB) energies (Wgt-2), A"-E" - the weight set with all the weights positive (Wgt-3), A, A', A" – correlation coefficients of the energy with TM-score over Cα atoms to the native structure after optimization with respect to the values before optimization, B, B', B" – distribution of correlation coefficients of the energy with Cα atom TM-score to the native structure before (open bars) and after (black bars) optimization of the force fields, C, C', C" – Z-score after optimization with respect to the values before optimization, D, D', D" – Cα atom TM-score to the native state of the lowest energy decoy after optimization with respect to the values before optimization, E, E', E" - Cα atom RMSD to the native state of the lowest energy decoy after optimization with respect to the values before optimization, open circles – results for the training protein set, black circles – results for the testing protein set.

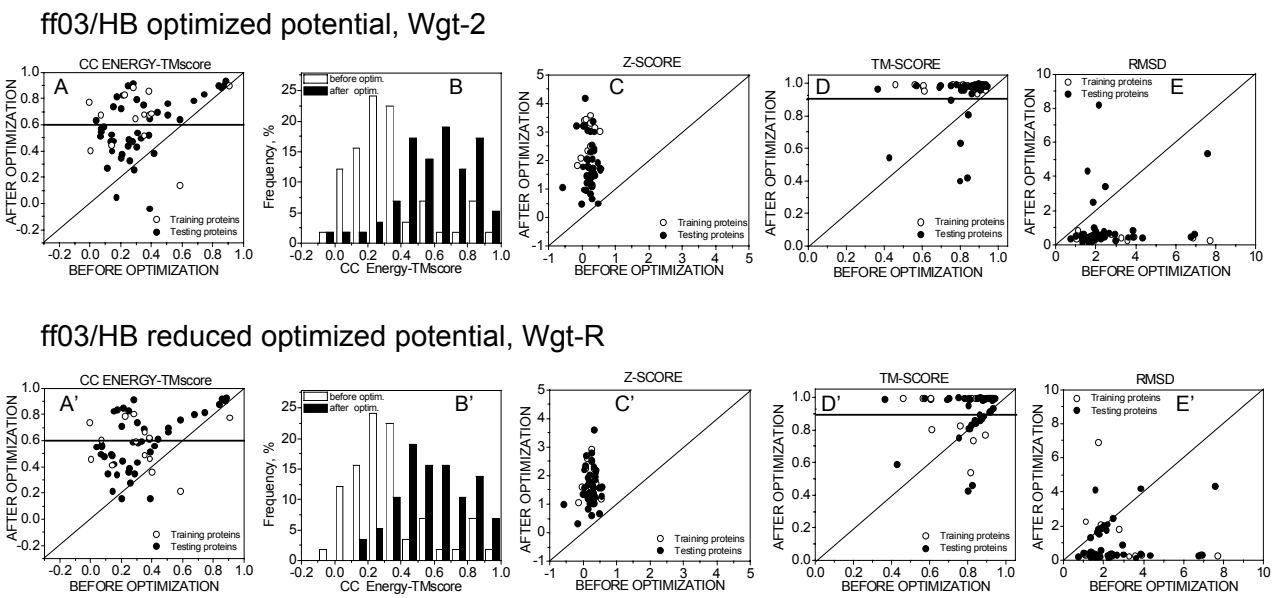


Figure C.4 (CHAPTER 3) Comparison of the scoring performance of the optimized ff03/HB, weight set Wgt-2 and the reduced optimized ff03/HB force fields, weight set Wgt-R, for the set of 58 proteins (Set58), A, A' – correlation coefficients of the energy with Cα atom TM-score to the native structure after optimization with respect to the values before optimization, B, B' – distribution of correlation coefficients of the energy with Cα atom TM-score to the native structure before (open bars) and after (black bars) optimization of the force fields, C, C' – Z-score after optimization with respect to the values before optimization, D, D' – Cα atom TM-score to the native state of the lowest energy decoy after optimization with respect to the values before optimization, E, E' – Cα atom RMSD to the native state of the lowest energy decoy after optimization with respect to the values before optimization, open circles – results for the training protein set, black circles – results for the testing protein set.

REFERENCES

1. Zhang, Y.; Arakaki, A. K.; Skolnick, J. Prot Struct Funct Bioinform 2005, 61, 91-98.
2. Baker, D.; Sali, A. Science 2001, 294, 93-96.
3. Ginalski, K.; Grishin, N. V.; Godzik, A.; Rychlewski, L. Nucleic Acids Res 2005, 33, 1874-1891.
4. Anfinsen, C. B. Science 1973, 181(4096), 223-230.
5. Bryngelson, J. D.; Wolynes, P. G. Proc Natl Acad Sci USA 1987, 84(21), 7524-7528.
6. Onuchic, J. N.; Wolynes, P. G. Curr Opin Struct Biol 2004, 14(1), 70-75.
7. Kim, P. S.; Baldwin, R. L. Annu Rev Biochem 1990, 59, 631-660.
8. Dill, K. A. Biochemistry 1990, 29(31), 7133-7155.
9. Baker, D.; Agard, D. A. Biochemistry 1994, 33, 7505-7509.
10. Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharan, V. J Mol Biol 1963, 7, 95-99.
11. Kolinski, A.; Skolnick, J. Proteins 1998, 32, 475-494.
12. Zhang, Y.; Kolinski, A.; Skolnick, J. Biophysical J 2003, 85, 1145-1164.
13. Zhang, Y.; Skolnick, J. PNAS 2004, 101(20), 7594-7599.
14. Kolinski, A.; Bujnicki, J. Proteins 2005, 61(Suppl 7), 84-90.
15. Bradley, P.; Misura, K. M. S.; Baker, D. Science 2005, 309, 1868-1871.
16. Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. J Am Chem Soc 1996, 118, 11225-11236.
17. Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S. J.; Weiner, P. K. J Am Chem Soc 1984, 106, 765-784.
18. Milik, M.; Kolinski, A.; Skolnick, J. J Comput Chem 1997, 18, 80-85.
19. Rotkiewicz, P.; Skolnick, J. J Comput Chem, (submitted).
20. Kirkpatrick, S.; Gelatt, C. D., Jr. ; Vecchi, M. P. Science 1983, 220, 671-680.
21. Li, Z.; Scheraga, H. A. Proc Natl Acad Sci USA 1987, 84(19), 6611–6615.
22. Lee, J.; Scheraga, H. A.; Rackovsky, S. J Comput Chem 1998, 18(9), 1222-1232.

23. Piela, L.; Kostrowicki, J.; Scheraga, H. A. J Phys Chem 1989, 93(8), 3339 - 3346.
24. Godzik, A.; Kolinski, A.; Skolnick, J. J Comput-Aided Mol Design 1993, 7, 397-438.
25. Godzik, A.; Kolinski, A.; Skolnick, J. J Comput Chem 1994, 14, 1194-1202.
26. Jagielska, A.; Skolnick, J. J Comp Chem 2007, 28(10), 1648-1657.
27. Sharp, K. A.; Honig, B. Annu Rev Biophys Biophys Chem 1990, 19, 301-332.
28. Tsui, V.; Case, D. A. Biopoymers (Nucl Acid Sci) 2001, 56, 275-291.
29. Onufriev, A.; Bashford, D.; Case, D. A. PROTEINS: Struct Funct Bioinf 2004, 55(2), 383-394.
30. Sitkoff, D.; Sharp, K. A.; Honig, B. J Phys Chem 1994, 98, 1978-1988.
31. Vorobjev, Y. N.; Almagro, J. C.; Hermans, J. PROTEINS: Struct Funct Bioinf 1998, 32(4), 399-413.
32. Moult, J.; Pedersen, J. T.; Judson, R.; Fidelis, K. PROTEINS: Struct Funct Gen 1995, 23(3), ii-iv.
33. Betz, S. F.; Baxter, S. M.; Fetrow, J. S. Drug Discov Today 2002, 7(16), 865-871.
34. Wallace, A. C.; Borkakoti, N.; Thornton, J. M. Protein Sci 1997, 6(11), 2308-2323.
35. Zhao, S.; Morris, G. M.; Olson, A. J.; Goodsell, D. S. J Mol Biol 2001, 314(5), 1245-1255.
36. Liang, M. P.; Brutlag, D. L.; Altman, R. B. Pac Symp Biocomput 2003, 204-215.
37. Fetrow, J. S.; Skolnick, J. J Mol Biol 1998, 281(5), 949-968.
38. Arakaki, A. K.; Zhang, Y.; Skolnick, J. Bioinformatics 2004, 20, 1087-1096.
39. Bissantz, C.; Bernard, P.; Hibert, M.; Rognan, D. Proteins 2003, 50(1), 5-25.
40. Clausen, H.; Buning, C.; Rarey, M.; Lengauer, T. J Mol Biol 2001, 308, 377-395.
41. Schonbrun, J.; Wedemeyer, W. J.; Baker, D. Curr Opin Struct Biol 2002, 12, 348-354.
42. Tramontano, A.; Morea, V. Proteins 2003, 53 Suppl 6, 352-368.
43. Duan, Y.; Kollman, P. A. Science 1998, 282, 740-744.
44. Pande, V. S.; Baker, I.; Chapman, J.; Elmer, S. P.; Khaliq, S.; Larson, S. M.; Rhee, Y. M.; Shirts, M. R.; Snow, C. D.; Sorin, E. J.; Zagrovic, B. Biopolymers 2003, 68, 91-109.

45. Lei, H.; Wu, C.; Liu, H.; Duan, Y. Proc Natl Acad Sci USA 2007, 104(12), 4925-4930.
46. Simmerling, C.; Strockbine, B.; Roitberg, A. E. J Am Chem Soc 2002, 124, 11258-11259.
47. Jagielska, A.; Scheraga, H. A. J Comput Chem 2007, 28(6), 1068-1082.
48. Simmerling, C.; Lee, M. R.; Oritz, A. R.; Kolinski, A.; Skolnick, J.; Kollman, P. A. J Am Chem Soc 2000, 122, 8392-8402.
49. Lee, M. R.; Tsai, J.; Baker, D.; Kollman, P. A. J Mol Biol 2001, 313, 417-430.
50. Lee, M. R.; Baker, D.; Kollman, P. A. J Am Chem Soc 2001, 123, 1040-1046.
51. Vieth, M.; Kolinski, A.; Brooks, C. L., III ; Skolnick, J. J Mol Biol 1994, 237, 361-367.
52. Lu, H.; Skolnick, J. Biopolymers 2003, 70, 575-584.
53. Fan, H.; Mark, A. E. Protein Sci 2004, 13, 211-220.
54. Jonassen, I.; Klose, D.; Taylor, W. R. Comput Biol Chem 2006, 30, 360-366.
55. Chen, J.; Brooks III, C. L. PROTEINS: Struct Funct Bioinf 2007, 67, 922-930.
56. Simons, K. T.; Bonneau, R.; Ruczinski, I.; Baker, D. Proteins 1999, Suppl 3, 171-176.
57. Lundstrom, J.; Rychlewski, L.; Bujnicki, J.; Elofsson, A. Protein Sci 2001, 10(11), 2354-2362.
58. Fisher, D. PROTEINS: Struct Funct Genet 2003, 51, 434-441.
59. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. Nucleic Acids Res 2000, 28(1), 235-242.
60. Zhang, Y.; Skolnick, J. Biophysical J 2004, 87, 2647-2655.
61. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. J Am Chem Soc 1995, 117, 5179-5197.
62. Case, D. A.; Cheatham, I., T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A., Jr.; Simmerling, C.; Wang, B.; Woods, R. J Comput Chem 2005, 26, 1668-1688.
63. Hsieh, M.-J.; Luo, R. PROTEINS: Struct Funct Bioinf 2004, 56, 475-486.
64. Lee, M. C.; Duan, Y. PROTEINS: Struct Funct Bioinf 2004, 55, 620-634.

65. Lazaridis, T.; Karplus, M. J Mol Biol 1998, 288, 477-487.
66. Dominy, B. N.; Brooks III, C. L. J Comput Chem 2002, 23, 147-160.
67. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. J Comput Chem 1983, 4, 187-217.
68. Felts, A. K.; Gallicchio, E.; Wallqvist, A.; Levy, R. M. PROTEINS: Struct Funct Genet 2002, 48, 404-422.
69. Summa, C. M.; Levitt, M.; DeGrado, W. F. J Mol Biol 2005, 352, 986-1001.
70. Zhou, H.; Zhou, Y. Protein Sci 2002, 11, 2714-2726.
71. Samudrala, R.; Xia, Y.; Huang, E.; Levitt, M. PROTEINS: Struct Funct Genet 1999, 37(Suppl. 3), 194-198.
72. Zhang, Y.; Skolnick, J. PROTEINS: Struct Funct Bioinf 2004, 57, 702-710.
73. James, F.; CERN Geneva, Switzerland: CERN Program Library Long Writeup D506, 1998.
74. Duan, Y.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. A. J Comp Chem 2003, 24(16), 1999-2012.
75. Sakae, Y.; Okamoto, Y. J Theor Comput Chem 2004, 3(3), 339-358.
76. Sakae, Y.; Okamoto, Y. J Theor Comput Chem 2004, 3(3), 359-378.
77. Mackerell, A. D.; Feig, M.; Brooks, C. L., III. J Comput Chem 2004, 25, 1400-1415.
78. Lwin, T. Z.; Luo, R. Protein Sci 2006, 15, 2642-2655.
79. Zhou, R. PROTEINS: Struct Funct Gen 2003, 53, 148-161.
80. Mongan, J.; Simmerling, C.; McCammon, J. A.; Case, D. A.; Onufriev, A. J Chem Theory Comput 2007, 3(1), 156-169.
81. Wagoner, J. A.; Baker, N. A. PNAS 2006, 103(22), 8331-8336.
82. Wroblewska, L.; Skolnick, J. J Comput Chem 2007, 28(12), 2059-2066.
83. Scheraga, H. A.; Liwo, A.; Oldziej, S.; Czaplewski, C.; Pillardy, J.; Ripoll, D. R.; Vila, J. A.; Kazmierkiewicz, R.; Saunders, J. A.; Arnautova, Y. A.; Jagielska, A.; Chinchio, M.; Nanias, M. Frontiers in Bioscience 2004, 9, 3296-3323 Suppl. S.
84. Oldziej, S.; Lagiewka, J.; Liwo, A.; Czaplewski, C.; Chinchio, M.; Nanias, M.; Scheraga, H. A. J Phys Chem B 2004, 108(43), 16950-16959

85. Arnautova, Y. A.; Jagielska, A.; Pillardy, J.; Scheraga, H. A. J Phys Chem B 2003, 107(29), 7143-7154.
86. Jagielska, A.; Arnautova, Y. A.; Scheraga, H. A. J Phys Chem B 2004, 108(32), 12181-12196.
87. Arnautova, Y. A.; Jagielska, A.; Scheraga, H. A. J Phys Chem B 2006, 110(10), 5025-5044.
88. Hansmann, U. H. E. Chem Phys Lett 1997, 281, 140-150.
89. Swendsen, R. H.; Wang, J. S. Phys Rev Lett 1986, 57, 2607-2609.
90. Zhang, Y.; Kihara, D.; Skolnick, J. PROTEINS: Struct Funct Gen 2002, 48, 192-201.
91. Betancourt, M. R. J Chem Phys 2005, 123, 174905.
92. Csendes, T. Acta Cybernetica 1988, 8, 361-370.
93. Yang, J. S.; Chen, W. W.; Skolnick, J.; Shakhnovich, E. I. Structure 2007, 15(1), 53-63.
94. Kabsch, W.; Sander, C. Biopolymers 1983, 22, 2577-2637.
95. Zhang, Y.; Hubner, I. A.; Arakaki, A. K.; Shakhnovich, E. I.; Skolnick, J. PNAS 2006, 103(8), 2605-2610.
96. Wroblewska, L.; Jagielska, A.; Skolnick, J. Biophys J 2007 (submitted).